联邦学习中隐私威胁及对策综述

Masahiro HAYASHITANI, Junki MORI, and Isamu TERANISHI
Secure System Platform Research Laboratories, NEC Corporation
hayashitani@nec.com, junki.mori@nec.com, teranisi@nec.com

摘要—联邦学习（FL）作为一种保护隐私的机器学习范式，能够在不直接交换客户端原始数据的情况下实现协作模型训练。虽然 FL 缓解了集中式数据收集相关的隐私风险，但它仍然容易受到各种可能泄露敏感信息的隐私威胁的影响。现有文献仅关注水平联邦学习（HFL）和垂直联邦学习（VFL）中的隐私威胁及对策。本文提供了对所有三种主要 FL 范式中隐私威胁的全面系统性回顾：HFL、VFL 以及联邦迁移学习（FTL）。我们引入了一个统一的分类法，根据目标数据类型对隐私威胁进行分类，并讨论了相应的防御机制。

***Index Terms*—水平联邦学习，垂直联邦学习，联邦迁移学习，隐私威胁，应对隐私威胁的对策。**

I. 介绍

随着计算设备的日益普及，个人每天生成大量的数据。这类数据的集中收集和存储非常耗费资源且耗时 [1]。此外，收集用户数据会引发严重的隐私和保密问题，因为这些数据通常包含敏感的个人信息。鉴于社会对隐私问题意识的增强，诸如欧盟人工智能法案等监管框架已经出现，进一步限制了大规模数据聚合实践的可能性。

针对这些挑战，联邦学习（FL）已成为一种保护隐私的机器学习范式，它使多个客户端能够在不共享其原始本地数据的情况下协作训练一个全局模型。FL 方法通常根据客户端之间数据特征和样本的一致性分为三大类：横向联邦学习（HFL）、纵向联邦学习（VFL）和联邦迁移学习（FTL）。

这些方法被广泛认为有利于保护隐私，因为它们减少了暴露个体级别原始数据的必要性。然而，FL 仍然容易受到隐私风险的影响。各种研究表明，在训练过程中传递模型更新可能会无意中向其他客户端、中央服务器或外部对手泄露敏感信息 [1]。

尽管针对特定形式的联邦学习（如 HFL 和 VFL）所固有的隐私风险已经在之前的综述研究中进行了考察和总结（例如，[1] 对于 HFL 和 [2] 对于 VFL），但仍缺乏对所有联邦学习范式（包括 FTL）进行统一和系统性回顾。本文全面概述了跨所有联邦学习变体的隐私风险，并根据所针对的数据性质将其分类为六种不同类型的隐私威胁。我们已经搜索了关于所有联邦学习范式中隐私威胁和对策的相关论文。该综述指导在每个联邦学习设置下需要考虑的隐私风险及其相应的缓解策略。

II. 联邦学习的分类

我们首先回顾由杨等人介绍的三种基于客户端间数据结构的 FL，即：[3]：HFL、VFL 和 FTL。图 1显示了每种类型在客户端之间的数据结构。HFL 假设每个客户端具有相同的功能和标签但样本不同（图 1(a)）。相比之下，VFL 假设客户端共享相同的样本但拥有不相交的特征（图 1(b)）。FTL 解决了客户端在样本和功能上都不同的情况（图 1(c)）。

以下子章节描述了每种联邦学习类型的学习和预测方法。

A. 水平联邦学习

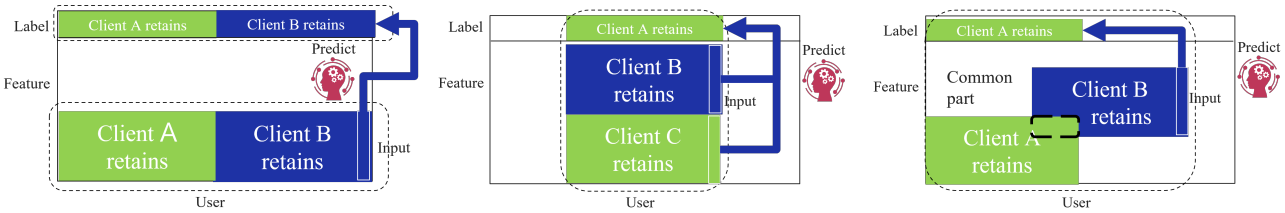
HFL 是谷歌首次引入的最常见联邦学习类别 [4]。HFL 的目标是每个客户端持有不同的样本，并协作改进具有共同结构模型的准确性。

图 2展示了 HFL 学习协议的概述。有两种实体参与 HFL 的学习：

i **服务器** - 协调者。服务器与客户端交换模型参数并聚合从客户端接收到的模型参数。

ii **客户** - 数据所有者。每个客户端使用自己的私有数据本地训练一个模型，并与服务器交换模型参数。

每个客户端首先训练一个局部模型几个步骤，并将模型参数发送到服务器。然后，服务器通过聚合局部模型来更新全局模型，通常通过平均，如 FedAvg 中所做的那样，并将结果分发给所有客户端。此过程重复直到收



(a) 水平联邦学习。 (b) 垂直联邦学习。 (c) 联邦迁移学习。

图 1. 基于客户端拥有的数据结构对联邦学习进行分类。

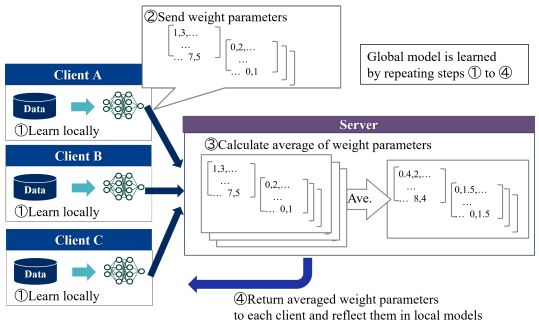


图 2. HFL 学习协议概述。

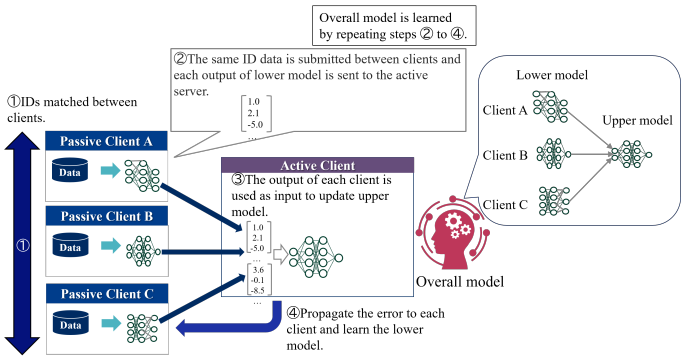


图 3. 标准 VFL 学习协议概述。

敛。在推理过程中，每个客户端使用全局模型及其特征独立预测标签。

该协议被称为集中式 HFL，因为它依赖于一个可信的第三方服务器。相比之下，去中心化的 HFL 消除了中央服务器，允许客户端直接通信，从而减少通信成本 [5]。

B. 垂直联邦学习

VFL 允许持有同一样本不同特征的客户端协同训练一个将每个客户端的各种特征作为输入的模型。有关 VFL 的研究涵盖了包括线性/逻辑回归 [6]–[10]、决策树 [11]–[15]、神经网络 [16]–[19] 和其他非线性模型 [20], [21] 在内的各种模型。

图 3展示了标准 VFL 学习协议的概述。在 VFL 中，只有单一客户端持有标签，并且它扮演服务器的角色。因此，参与 VFL 学习的有两种类型的实体：

i **活跃客户端** - 特征和标签所有者。活跃客户端协调学习过程。它计算损失并与被动客户端交换中间输出。

ii **被动客户端** - 特征所有者。每个被动客户端都将其特征和模型保持本地，但与主动客户端交换中间输出。

VFL 包含两个阶段：ID 匹配阶段和学习阶段。在 ID 匹配阶段，所有客户端共享公共样本 ID。在学习阶段，每个客户端都有一个单独的模型，并将其自身的特征作为输入，被动客户端将计算出的中间输出发送给主动客户端。主动客户端根据聚合的中间输出计算损失，并将梯度发送给所有被动客户端。然后，被动客户端更新自己的模型参数。这个过程重复进行直到收敛。在推理阶段，所有客户端需要合作来预测样本的标签。

C. 联邦迁移学习

FTL 假设两个仅共享一小部分样本或特征的客户端。FTL 目的是通过将拥有标签的客户端（源客户端）的知识转移到没有标签的客户端（目标客户端），为目标客户端创建一个可以预测标签的模型。

图 4展示了 FTL 学习协议的总体情况。如上所述，有两种类型的实体参与 FTL：

i **源客户端** - 特征和标签所有者。源客户端与目标客户端交换中间输出，如输出和梯度，并计算损失。

ii **目标客户** - 特性所有者。目标客户端与源客户端交换中间输出。

在 FTL 中，两个客户端交换中间输出以学习共同的表示。源客户端使用标记数据计算损失并将梯度发送给目标客户端，后者更新目标客户端的表示。该过程重复进行直到收敛。在推理过程中，目标客户端使用其自己的模型和特征预测样本的标签。

学习协议的细节因具体方法而异。尽管提出的 FTL 方法数量有限，但我们介绍了三种重要的方法。FTL 需要一些补充信息来连接两个客户端，例如公共 ID [22]–[25]、公共特征 [26], [27] 以及目标客户端的标签 [28], [29]。

1) ID-FTL: 大多数 FTL 方法假设两个客户端之间存在共同 ID 的样本。如同 VFL，这种类型的 FTL 在学习阶段之前需要进行 ID 匹配。刘等人。[22] 提出了第一个 FTL 协议，该协议学习特征转换函数，使得不同样本的特征被映射到相同的特征上。夏尔玛等人。[23] 通过多方计算改进了首个 FTL 的通信开销，并通过处理恶意客户端增强了安全性。高等人。[25] 提出了一种双学习框架，在此框架中，两个客户端通过交换他们在共享样本上的插补模型输出来互相补充对方缺失的特征。

2) 特征-FTL: 在实际应用中，共享具有相同 ID 的样本是困难的。因此，高等人。[26] 提出了通过假设共同特征而不是共同样本来实现 FTL 的方法。在这种方法中，两个客户端使用交换的特征映射模型相互重建缺失的特征。然后，利用所有特征，客户端进行 HFL 以获得标签预测模型。在原论文中，作者假定所有客户端都拥有标签。然而，由于源客户端只能自行学习标签预测模型，这种方法适用于没有标签的目标客户端。森等人。[27] 提出了神经网络的方法，在该方法中每个客户端除了共同特征外还将其自身的独特特征纳入 HFL 训练。然而，他们的方法基于 HFL，不能应用于不具有标签的目标客户端。

3) 标签-FTL: 该方法假设既没有共同的 ID 也没有特征，但假定所有客户端都拥有标签，从而可以在客户端之间学习到一个通用表示。由于它是基于 HFL 的，因此参与实体与 HFL 中的相同。高等人。[28] 通过与服务器交换中间输出并减少最大均值差异损失来学习一个共同表示。拉科托莫纳奇等人。[29] 提出了一种使用 Wasserstein 距离对中间输出进行处理的方法，以学习一个通用表示，从而使客户端仅能与服务器交换诸如平均值和方差之类的统计信息。

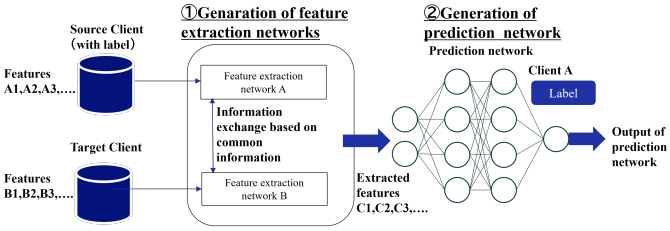


图 4. 整体的 FTL 学习协议。

III. 威胁模型和联合学习中的隐私攻击

本节展示了威胁模型并分类了隐私攻击。

A. 威胁模型

我们基于四个视角定义威胁模型：FL 类型、攻击者角色、攻击信息和攻击风格。

1) 攻击者角色: 我们定义三种攻击者角色：服务器、客户端和第三方。

i **服务器** - 表示由服务器发起的攻击。在 VFL 和 FTL 中，活跃客户端和源客户端分别对应这一角色。

ii **客户端** - 表示由客户端发起的攻击。在 VFL 和 FTL 中，这分别对应于被动客户端和目标客户端。

iii **第三方** - 指代既不属于服务器也不属于客户端角色的攻击者。

2) 攻击信息: 攻击者可以利用共享信息，如梯度或模型参数，来推断私有数据。在训练过程中，即使只共享一小部分信息，交换梯度也可能暴露敏感信息并导致深度泄露。这些梯度能够揭示关于本地数据的重要细节。从共享的本地模型参数中，攻击者可能推断出类别代表、数据集成员身份和特征，甚至重构原始训练输入和标签。攻击者可以访问的具体信息取决于他们在系统中的角色：

i **服务器** - 每轮从每个客户端接收梯度或模型参数。

ii **客户端** - 每轮仅访问模型参数。

iii **第三方** - 可能窃听通信并访问梯度或模型参数。

3) 攻击风格: 我们定义了三种攻击风格：恶意、诚实但好奇和诚实。

i **恶意的** - 主动干扰训练过程以从客户端提取私有信息。

ii **诚实但好奇** - 被动跟随 FL 协议而不干扰训练，但试图从接收到的数据中推断私有信息。

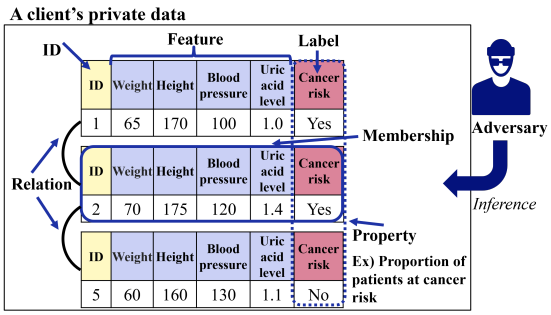


图 5. 隐私攻击图。

iii **诚实** - 被动遵循 FL 协议，并不试图从接收到的数据中推断私有信息。

如前一节所述，我们考虑三种类型的联邦学习：HFL、VFL 和 FTL。

4) 联合学习类型:

B. 隐私攻击

我们根据目标数据的性质将隐私攻击分为六类：特征推断、属性推断、成员推断、标签推断、ID 泄露和关系泄露。图 5阐述了这些类别的图表，表 I根据相应的威胁模型组织它们，并附有解决每种威胁的相关参考文献。虽然现有研究并未专门针对联邦学习（FTL）环境中的隐私威胁进行识别，但我们将“FTL”包含在“FL 类型”列中，用于那些可能适用于联邦学习的隐私威胁。以下描述每种威胁类型。

1) 特征推断: 特征推理攻击旨在恢复输入特征。如果所有特征都是目标，该攻击也称为重构攻击；如果仅针对子集，则称为属性推断。此类攻击首次由 [30] 为 HFL 引入。最近，各种攻击方法已经为 VFL 开发出来。这些攻击在使用梯度提升决策树 [31] 的 HFL 设置中尤其有效。基于 HFL 的攻击适用于 Feature-FTL 和 Label-FTL，它们在后期阶段包含 HFL，而基于 VFL 的攻击对 ID-FTL 有效，其中使用了通用的标识符。

2) 属性推断: 属性推断攻击针对本地数据集的属性，如类别分布 [32]、敏感属性 [33] 或与主要任务无关的属性 [34]。这些攻击在 HFL 中相关，因为客户端的数据分布可能不同。它们通常假设对手拥有带有目标属性标签的辅助训练数据 [34]。

3) 成员推断: 成员推断攻击试图确定特定数据点是否为客户训练集的一部分。例如，攻击者可能推测某一特定患者档案是否被用于训练疾病分类器。这些攻击在 HFL 中特别有效，因为每个客户都有独特数据样本且服务器有权访问模型参数，从而能够进行强大的白盒攻击 [35]。

4) 标签推理: 标签推断攻击旨在在某些客户端缺乏标签信息时推断样本的标签。在 VFL 中，被动客户端不拥有标签，但交换中间计算结果，从而可能导致标签泄露。VFL 架构天生易受此类攻击影响 [36]。这些攻击同样有效针对 ID-FTL 场景，在该场景下，客户端共享样本，但目标客户端没有标签。

5) ID 泄漏: 在 VFL 和 ID-FTL 中，这些方法要求跨客户端对共享样本进行对齐，存在一个风险，即一个客户端持有的样本 ID 可能会暴露给其他客户端。因此，即使是诚实的客户端也可能获得有关实体存在的知识，这些实体要么不属于交集 [37]，要么位于共享样本的交集中 [38]。当客户端之间的样本 ID 不对称时，其中一个客户端的所有 ID 集合都会被另一个客户端暴露出来 [39]，这带来了重大的隐私风险。这种固有的脆弱性严重限制了 VFL 和 ID-FTL 的实际应用。

6) 关系泄露: 在应用于关系数据（例如图）的 FL 中，可能会发生关系泄漏，暴露样本之间的连接 [40]。这些泄漏利用了相关样本在嵌入空间中的表示通常接近这一直觉。如果两个表示之间的距离低于某个阈值，则可以推断出一种关系。

IV. 一般防御方法

我们总结了上一节中描述的针对隐私攻击的防御方法。这些方法主要分为两大类：通用方法，对大多数攻击都有效；以及专门方法，针对特定攻击但更高效。虽然专门方法是为特定威胁量身定制的，但它们通常比通用方法更简单且计算成本更低，因此非常实用。本节特别关注通用防御方法。

在保护隐私的联邦学习中广泛使用的两种通用方法是：（1）通信渠道防御，它采用加密技术来保护模型参数或梯度；（2）差分隐私（DP），它防止数据记录从模型中泄露。

A. 通信信道防御

这些防御措施通过加密服务器和客户端之间共享的原始中间输出（如局部模型参数、梯度和输出），防止信息泄露。然而，它们会带来高昂的计算成本。我们回顾了该类别中的两种主要技术：安全多方计算 (SMPC) 和区块链。

表 I
隐私威胁和威胁模型

威胁类型	FL 类型	参考文献	攻击者角色	攻击信息	攻击风格
特征推断	HFL (Feature-FTL, Label-FTL)	[41], [42], [43], [44], [45]	Server	Gradient	Malicious
		[30], [46], [47], [48] [49], [50], [51], [52]	服务器	梯度	诚实但好奇
		[53], [54]	Server	Model Parameter	Malicious
		[55], [56]	Client	Model Parameter	Malicious
		[57]	3rd party	Model Parameter	Malicious
		[58]	3rd party	Gradient	Honest-but-curious
		[59], [60], [61]	Client	Model Parameter	Honest-but-curious
	VFL（ID-FTL）	[62], [63], [64]	Server	Gradient	Malicious
		[65]	Server	Gradient	Honest-but-curious
		[66]	Server	Model Parameter	Honest-but-curious
		[67], [68], [69]	Client	Model Parameter	Honest-but-curious
属性推断	HFL	[32], [70]	Server	Gradient	Malicious
		[71]	Server	Model Parameter	Malicious
		[34], [72], [73]	Client	Model Parameter	Malicious
		[33], [72]	Client	Model Parameter	Honest-but-curious
成员推断	HFL	[35]	Server	Gradient	Malicious
		[35]	Server	Gradient	Honest-but-curious
		[74], [35], [75], [76]	Client	Model Parameter	Malicious
		[35], [77]	Client	Model Parameter	Honest-but-curious
标签推理	VFL（ID-FTL）	[36], [78]	Client	Model Parameter	Malicious
		[79], [80], [81], [82]	Client	Model Parameter	Honest-but-curious
ID 泄漏	VFL（ID-FTL）	[37], [39]	Server, Client	Non-Aligned IDs	Honest
		[38]	Server, Client	Aligned IDs	Honest
Relation Leaks	VFL (ID-FTL)	[40]	Client	Model Parameter	Honest-but-curious

1) 安全多方计算（SMPC）： SMPC [83] 允许多方在不向彼此泄露输入的情况下联合计算函数。SMPC 确保任何一方都无法获得最终输出之外的任何信息。SMPC 天然适合用于安全聚合 HFL 中的本地模型参数和梯度，特别是在服务器不可信（恶意或诚实但好奇）时。在 VFL 和 FTL 中，由于必须共享更多敏感的中间输出，因此通常将 SMPC 技术集成到大多数提出的算法中。然而，尽管 SMPC 保护了中间输出，但它并不能保障最终聚合的输出的安全性，后者仍然容易受到推理攻击。实现 SMPC 的三种主要方法是：同态加密（同态加密 HE）、乱线电路（Garbled Circuits GC）和秘密共享（秘密共享 SS）。

同态加密（HE）。是一种加密技术，它允许在不解密的情况下对加密输入进行诸如加法和乘法等计算。例如，作为密码学技术，Paillier [84] 和修改后的 El Gamal [85] 具有加法同态性质，而 El Gamal 和 RSA [86] 具有乘法同态性质。

乱线电路（GC）。GC，由 Yao 首次提出 [87]，构建布尔电路用于安全的两方计算。关于 GC 的研究主要集中在提高性能和加强安全性上。安全性的进步解决了来自半诚实对手和恶意对手的威胁。同时，在不牺牲安全性的前提下优化 GC 性能的努力仍然是一个活跃的研究领域。

秘密共享（SS）。分发秘密共享是一种加密技术，它将一个秘密分散到多个参与者中，因此没有任何单个个体持有完整的秘密。只有当足够的份额被组合时，原始的秘密才能被重建。秘密共享使在不向任何参与者暴露原始数据的情况下进行机器学习模型的安全训练成为可能。该模型被分成若干份并分发出去，仅在收集足够多的份额时才允许重建。

2) 区块链： 区块链是一个分布式账本，它使用加密哈希安全地链接不断扩大的记录列表（区块）。它支撑着大多数数字货币，并使点对点联邦学习成为可能，而无需依赖潜在不可信的服务器，这与 SMPC 和 HE 不同。

B. 差分隐私（DP）

FL 协议通常会结合额外的 DP 技术，以防止信息从本地更新和全局模型中泄露。DP 确保任何个人数据的包含或排除对输出的影响有限，从而保护敏感信息。这种保护是通过向输出添加校准过的随机噪声来实现的，该噪声与单个数据点可能产生的最大影响成正比。重要的是，DP 假设对手可能拥有任意的外部知识。存在几种将 DP 集成到 FL 中的方法，以保障客户端数据隐私。

1) 中心差分隐私（CDP）： CDP 是原始的 DP 定义。以下是 DP 的定义。

定义 1: 对于 $\epsilon > 0$ 和 $0 \leq \delta < 1$，一个随机机制 $\mathcal{M}$ 被称为 (ϵ, δ)-差分隐私当且仅当对所有 $S \subseteq \text{Range}(\mathcal{M})$ 以及所有相邻数据集 D 和 D'，我们有

$$\Pr[\mathcal{M}(\mathrm{D}) \in \mathrm{S}] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(\mathrm{D}') \in \mathrm{S}] + \delta$$

在 FL 中，CDP 通过一个可信服务器实现，该服务器向聚合输出（如全局模型或梯度）添加噪声。这种噪声降低了单个客户端推断有关其他客户端信息的能力。然而，添加噪声通常会导致性能下降。此外，CDP 对不可信服务器无效，因为它依赖于服务器来应用噪声。

2) 局部差分隐私（LDP）： 我们可以将差分隐私扩展到服务器不可信的情况，每个客户端在其自身的中间输出中添加噪声并将其与服务器共享。这是通过局部差分隐私实现的。

LDP 确保每个客户端的隐私受到其他客户端和服务器的保护。然而，LDP 要求每个客户端添加足够的噪声，因此添加的总噪声量巨大，导致性能下降比 DP 更为严重。

3) 分布式差分隐私（DDP）： DDP 通过集成加密协议来保护个人隐私，弥合了 LDP 与 CDP 之间的差距。DDP 避免依赖可信服务器，并且比 LDP 实现了更好的效用。理论上，由于两者添加的噪声量相同，DDP 在效用上与 CDP 相匹配。

DDP 反映了添加到统计量中的噪声在客户端之间进行了分配。实现通常涉及每个参与者应用具有减小方差的噪声机制。这些机制需要稳定的分布以确保整体输出经过良好校准，并且需要加密技术来隐藏中间输出不被客户端看到。

4) 参与者级别的差分隐私（PDP）： 上述的动态规划技术可以防止个别客户端数据的泄露，但对于针对每个客户端数据属性的推断攻击却无效。需要使用 PDP 机制来解决这一限制，该机制会在客户端级别模糊信息。PDP 通过协调并添加噪声使各个客户端不可区分来实现这一点。

V. 专用防御方法

这些防御方法针对特定攻击，但比一般方法更简单且计算强度更低，因此非常实用。

A. 特征推断

使用较大的批量大小是一种潜在的防御隐私攻击的方法 [46], [59]。更大的批次通过引入更多变量增加了优化复杂性。采用线性规划方法从梯度中重构全批数据 [59]。然而，随着约束条件数量的增加，解的时间显著增长，使得此类攻击在计算上不可行。

另一种简单的防御方法是 dropout，这是一种正则化技术，通过随机失活神经元来减少过拟合 [32], [34]。Dropout 通过减少对手可见的活跃梯度数量来削弱攻击。这种随机特征移除进一步混淆了训练期间的敏感数据。

B. 属性推断

对于属性推断，dropout 再次成为一种可行的防御措施 [32], [34]。另一种方法是共享较少的梯度更新，其中参与者每轮只披露他们的一部分梯度 [34]。这既减少了通信成本也降低了潜在的信息泄露。

C. 成员推断

在集中式机器学习中有效的缓解技术也可以应用于 FL 中的本地训练。例如，常用以防止过拟合的正则化、dropout 和蒸馏方法也显示出了对成员推断攻击的有效性。此外，在客户端训练中集成经过实证验证的安全措施，这些安全措施旨在提高对白盒成员推断攻击的鲁棒性 [88], [89]，可以进一步降低隐私风险。

D. 标签推理

为减轻 VFL 中的标签推断威胁，提出了多种防御策略，这些策略主要集中在模糊梯度或破坏共享表示与私有标签之间的相关性上。诸如梯度扰动 [82] 和合成梯度生成 [79] 等技术已被证明在性能损失极小的情况下非常有效。基于优化的方法也通过最小化中间嵌入与私有标签之间的相关性来降低对手的推断能力 [78], [80]。此外，还出现了特定架构的防御措施，例如结合标签

差分隐私和后处理，并将互信息正则化应用于树模型 [81]。

E. ID 泄漏

私有集合交集（PSI）是 VFL 和 ID-FTL 中用于防止在 ID 对齐过程中泄露非重叠 ID 信息的标准加密技术 [37]。为了支持不对称的 ID 对齐，已经引入了一种改进的 PSI 变体 [39]。然而，这些技术仍然会揭示交集内的成员信息，从而破坏完全隐私。为了解决这一问题，提出了插入虚拟 ID 的方法 [38]，使客户端可以在不知道哪些 ID 匹配的情况下继续操作，并增强共享 ID 的隐私。

VI. 结论

此次调查显示，联邦学习仍然容易受到多种特定范式的隐私威胁。我们的分类法表明，尽管 HFL 已经得到了广泛的研究，特别是在特征和成员推断方面，但针对属性推断（如参与者级别的差分隐私）的防御措施仍然不成熟。相比之下，VFL 面临着固有的风险，包括标签推断和 ID 泄露，这需要更加针对性的对策。根据设置的不同，FTL 继承了 HFL 和 VFL 的脆弱性，但是关于 FTL 特定风险及其防御的系统研究尚显不足。未来的研究应该优先考虑减轻 VFL 的内在脆弱性，对 FTL 特有的威胁进行深入分析，并推进统一的威胁模型和基准框架的发展。最后，开发轻量级、通用的防御措施，在应对多种威胁的同时保持其效用，对于实现联邦学习的实际部署及安全性至关重要。

致谢

这项研发包括了日本国家内部事务与通信省 (MIC) 的“通过安全数据协调优化 AI 技术的研究与发展 (JPMI00316)”的结果。

参考文献

[1] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, “Privacy and robustness in federated learning: Attacks and defenses,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, pp. 8726–8746, 2024.

[2] Y. Liu, Y. Kang, T. Zou, Y. Pu, Y. He, X. Ye, Y. Ouyang, Y.-Q. Zhang, and Q. Yang, “Vertical federated learning: Concepts, advances, and challenges,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3615–3634, 2024.

[3] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, 2019.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 54. PMLR, 2017, pp. 1273–1282.

[5] E. T. Martínez Beltrán, M. Q. Pérez, P. M. S. Sánchez, S. L. Bernal, G. Bovet, M. G. Pérez, G. M. Pérez, and A. H. Celdrán, “Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges,” *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2983–3013, 2023.

[6] A. Gascón, P. Schoppmann, B. Balle, M. Raykova, J. Doerner, S. Zahur, and D. Evans, “Secure linear regression on vertically partitioned datasets,” *IACR Cryptol. ePrint Arch.*, vol. 2016, p. 892, 2016.

[7] S. Hardy, W. Henecka, H. Ivey-Law, R. Nock, G. Patrini, G. Smith, and B. Thorne, “Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption,” *CoRR*, vol. abs/1711.10677, 2017.

[8] R. Nock, S. Hardy, W. Henecka, H. Ivey-Law, G. Patrini, G. Smith, and B. Thorne, “Entity resolution and federated learning get a federated resolution,” *CoRR*, vol. abs/1803.04035, 2018.

[9] S. Yang, B. Ren, X. Zhou, and L. Liu, “Parallel distributed logistic regression for vertical federated learning without third-party coordinator,” *CoRR*, vol. abs/1911.09824, 2019.

[10] Q. Zhang, B. Gu, C. Deng, and H. Huang, “Secure bilevel asynchronous vertical federated learning with backward updating,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 10 896–10 904, May 2021.

[11] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, D. Papadopoulos, and Q. Yang, “Secureboost: A lossless federated learning framework,” *IEEE Intelligent Systems*, vol. 36, no. 6, pp. 87–98, 2021.

[12] J. Vaidya, C. Clifton, M. Kantarcioglu, and A. S. Patterson, “Privacy-preserving decision trees over vertically partitioned data,” *ACM Trans. Knowl. Discov. Data*, vol. 2, no. 3, oct 2008.

[13] Y. Wu, S. Cai, X. Xiao, G. Chen, and B. C. Ooi, “Privacy preserving vertical federated learning for tree-based models,” *Proc. VLDB Endow.*, vol. 13, no. 12, p. 2090 – 2103, jul 2020.

[14] Y. Liu, Y. Liu, Z. Liu, Y. Liang, C. Meng, J. Zhang, and Y. Zheng, “Federated forest,” *IEEE Transactions on Big Data*, pp. 1–1, 2020.

[15] Z. Tian, R. Zhang, X. Hou, J. Liu, and K. Ren, “Federboost: Private federated learning for GBDT,” *CoRR*, vol. abs/2011.02796, 2020.

[16] Y. Hu, D. Niu, J. Yang, and S. Zhou, “Fdml: A collaborative machine learning framework for distributed features,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2232 – 2240.

[17] Y. Liu, Y. Kang, X. Zhang, L. Li, Y. Cheng, T. Chen, M. Hong, and Q. Yang, “A communication efficient collaborative learning framework for distributed features,” *CoRR*, vol. abs/1912.11187, 2019.

[18] D. Romanini, A. J. Hall, P. Papadopoulos, T. Titcombe, A. Ismail, T. Cebere, R. Sandmann, R. Roehm, and M. A. Hoeh, “Pyvertical: A vertical federated learning framework for multi-headed splitnn,” *CoRR*, vol. abs/2104.00489, 2021.

[19] Q. He, W. Yang, B. Chen, Y. Geng, and L. Huang, "Transnet: Training privacy-preserving neural network over transformed layer," *Proc. VLDB Endow.*, vol. 13, no. 12, p. 1849 – 1862, jul 2020.

[20] B. Gu, Z. Dang, X. Li, and H. Huang, "Federated doubly stochastic kernel learning for vertically partitioned data," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery, 2020, p. 2483 – 2493.

[21] R. Xu, N. Baracaldo, Y. Zhou, A. Anwar, J. Joshi, and H. Ludwig, "Fedv: Privacy-preserving federated learning over vertically partitioned data," *CoRR*, vol. abs/2103.03918, 2021.

[22] Y. Liu, Y. Kang, C. Xing, T. Chen, and Q. Yang, "A secure federated transfer learning framework," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 70–82, 2020.

[23] S. Sharma, C. Xing, Y. Liu, and Y. Kang, "Secure and efficient federated transfer learning," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 2569–2576.

[24] B. Zhang, C. Chen, and L. Wang, "Privacy-preserving transfer learning via secure maximum mean discrepancy," *arXiv preprint arXiv:2009.11680*, 2020.

[25] Y. Gao, M. Gong, Y. Xie, A. K. Qin, K. Pan, and Y.-S. Ong, "Multiparty dual learning," *IEEE Transactions on Cybernetics*, vol. 53, no. 5, pp. 2955–2968, 2023.

[26] D. Gao, Y. Liu, A. Huang, C. Ju, H. Yu, and Q. Yang, "Privacy-preserving heterogeneous federated transfer learning," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 2552–2559.

[27] J. Mori, I. Teranishi, and R. Furukawa, "Continual horizontal federated learning for heterogeneous data," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8.

[28] D. Gao, C. Ju, X. Wei, Y. Liu, T. Chen, and Q. Yang, "Hhhfl: Hierarchical heterogeneous horizontal federated learning for electroencephalography," *arXiv preprint arXiv:1909.05784*, 2019.

[29] A. Rakotomamonjy, M. Vono, H. J. M. Ruiz, and L. Ralaivola, "Personalised federated learning on heterogeneous feature spaces," *arXiv preprint arXiv:2301.11447*, 2023.

[30] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf

[31] K. Ito, B. Enkhtaivan, I. Teranishi, and J. Sakuma, "Trojan attribute inference attack on gradient boosting decision trees," in *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*, 2024, pp. 542–559.

[32] L. Wang, S. Xu, X. Wang, and Q. Zhu, "Eavesdrop the composition proportion of training labels in federated learning," 2019. [Online]. Available: https://arxiv.org/abs/1910.06044

[33] W. Zhang, S. Tople, and O. Ohrimenko, "Leakage of dataset properties in multi-party machine learning," 2021. [Online]. Available: https://arxiv.org/abs/2006.07267

[34] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," 2018. [Online]. Available: https://arxiv.org/abs/1805.04049

[35] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, May 2019. [Online]. Available: http://dx.doi.org/10.1109/SP.2019.00065

[36] C. Fu, X. Zhang, S. Ji, J. Chen, J. Wu, S. Guo, J. Zhou, A. X. Liu, and T. Wang, "Label inference attacks against vertical federated learning," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 1397–1414. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/fu-chong

[37] S. Hardy, W. Henecka, H. Ivey-Law, R. Nock, G. Patrini, G. Smith, and B. Thorne, "Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption," *arXiv preprint arXiv:1711.10677*, 2017.

[38] J. Sun, X. Yang, Y. Yao, A. Zhang, W. Gao, J. Xie, and C. Wang, "Vertical federated learning without revealing intersection membership," 2021. [Online]. Available: https://arxiv.org/abs/2106.05508

[39] Y. Liu, X. Zhang, and L. Wang, "Asymmetrical vertical federated learning," *arXiv preprint arXiv:2004.07427*, 2020.

[40] P. Qiu, X. Zhang, S. Ji, T. Du, Y. Pu, J. Zhou, and T. Wang, "Your labels are selling you out: Relation leaks in vertical federated learning," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 5, pp. 3653–3668, 2023.

[41] H. Ren, J. Deng, and X. Xie, "Grnn: Generative regression neural network—a data leakage attack for federated learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 4, may 2022. [Online]. Available: https://doi.org/10.1145/3510032

[42] J. Sun, A. Li, B. Wang, H. Yang, H. Li, and Y. Chen, "Soteria: Provable defense against privacy leakage in federated learning from representation perspective," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9311–9319.

[43] F. Boenisch, A. Dziedzic, R. Schuster, A. S. Shamsabadi, I. Shumailov, and N. Papernot, "When the curious abandon honesty: Federated learning is not private," 2023. [Online]. Available: https://arxiv.org/abs/2112.02918

[44] W. Wei, L. Liu, Y. Wu, G. Su, and A. Iyengar, "Gradient-leakage resilient federated learning," 2021. [Online]. Available: https://arxiv.org/abs/2107.01154

[45] M. Lam, G.-Y. Wei, D. Brooks, V. J. Reddi, and M. Mitzenmacher, "Gradient disaggregation: Breaking privacy in federated learning by reconstructing the user participant matrix," 2021. [Online]. Available: https://arxiv.org/abs/2106.06089

[46] B. Zhao, K. R. Mopuri, and H. Bilen, "idlg: Improved deep leakage from gradients," 2020. [Online]. Available: https://arxiv.org/abs/2001.02610

[47] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients - how easy is it to break privacy in federated learning?" in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[48] W. Wei, L. Liu, M. Loper, K.-H. Chow, M. E. Gursoy, S. Truex, and Y. Wu, "A framework for evaluating gradient leakage attacks in federated learning," 2020. [Online]. Available: https://arxiv.org/abs/2004.10397

[49] M. Vero, M. Balunović, D. I. Dimitrov, and M. Vechev, "Tableak: Tabular data leakage in federated learning," 2023. [Online]. Available: https://arxiv.org/abs/2210.01785

[50] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2024.

[51] A. Hatamizadeh, H. Yin, P. Molchanov, A. Myronenko, W. Li, P. Dogra, A. Feng, M. G. Flores, J. Kautz, D. Xu, and H. R. Roth, "Do gradient inversion attacks make federated learning unsafe?" *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, pp. 2044–2056, 2023.

[52] H. Liu, B. Li, C. Gao, P. Xie, and C. Zhao, "Privacy-encoded federated learning against gradient-based data reconstruction attacks," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 5860–5875, 2023.

[53] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," 2018. [Online]. Available: https://arxiv.org/abs/1812.00535

[54] M. Song, Z. Wang, Z. Zhang, Y. Song, Q. Wang, J. Ren, and H. Qi, "Analyzing user-level privacy attack against federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2430–2444, 2020.

[55] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: Information leakage from collaborative deep learning," 2017. [Online]. Available: https://arxiv.org/abs/1702.07464

[56] X. Xu, J. Wu, M. Yang, T. Luo, X. Duan, W. Li, Y. Wu, and B. Wu, "Information leakage by model weights on federated learning," in *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, ser. PPMLP'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 31 – 36. [Online]. Available: https://doi.org/10.1145/3411501.3419423

[57] X. Yuan, X. Ma, L. Zhang, Y. Fang, and D. Wu, "Beyond class-level privacy leakage: Breaking record-level privacy in federated learning," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2555–2565, 2022.

[58] H. Liang, Y. Li, C. Zhang, X. Liu, and L. Zhu, "Egia: An external gradient inversion attack in federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4984–4995, 2023.

[59] Z. Li, Z. Huang, C. Chen, and C. Hong, "Quantification of the leakage in federated learning," 2020. [Online]. Available: https://arxiv.org/abs/1910.05467

[60] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2019. [Online]. Available: https://arxiv.org/abs/1812.00984

[61] J. So, R. E. Ali, B. Güler, J. Jiao, and A. S. Avestimehr, "Securing secure aggregation: Mitigating multi-round privacy leakage in federated learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, pp. 9864–9873, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26177

[62] J. Sun, Y. Yao, W. Gao, J. Xie, and C. Wang, "Defending against reconstruction attack in vertical federated learning," 2021. [Online]. Available: https://arxiv.org/abs/2107.09898

[63] M. N. Vu, T. R. Jeter, R. Alharbi, and M. T. Thai, "Active data reconstruction attacks in vertical federated learning," in *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 1374–1379.

[64] X. Jiang, X. Zhou, and J. Grossklags, "Comprehensive analysis of privacy leakage in vertical federated learning during prediction," *Proceedings on Privacy Enhancing Technologies*, 2022.

[65] X. Jin, P.-Y. Chen, C.-Y. Hsu, C.-M. Yu, and T. Chen, "Cafe: catastrophic data leakage in vertical federated learning," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2024.

[66] X. Luo, Y. Wu, X. Xiao, and B. C. Ooi, "Feature inference attack on model predictions in vertical federated learning," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, Apr. 2021. [Online]. Available: http://dx.doi.org/10.1109/ICDE51399.2021.00023

[67] H. Weng, J. Zhang, X. Ma, F. Xue, T. Wei, S. Ji, and Z. Zong, "Practical privacy attacks on vertical federated learning," 2022. [Online]. Available: https://arxiv.org/abs/2011.09290

[68] F. Fu, H. Xue, Y. Cheng, Y. Tao, and B. Cui, "Blindfl: Vertical federated machine learning without peeking into your data," in *Proceedings of the 2022 International Conference on Management of Data*, ser. SIGMOD/PODS ' 22. ACM, Jun. 2022. [Online]. Available: http://dx.doi.org/10.1145/3514221.3526127

[69] R. Yang, J. Ma, J. Zhang, S. Kumari, S. Kumar, and J. J. P. C. Rodrigues, "Practical feature inference attack in vertical federated learning during prediction in artificial internet of things," *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 5–16, 2024.

[70] F. Mo, A. Borovykh, M. Malekzadeh, H. Haddadi, and S. Demetriou, "Layer-wise characterization of latent information leakage in federated learning," 2021. [Online]. Available: https://arxiv.org/abs/2010.08762

[71] M. Shen, H. Wang, B. Zhang, L. Zhu, K. Xu, Q. Li, and X. Du, "Exploiting unintended property leakage in blockchain-assisted federated learning for intelligent edge computing," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2265–2275, 2021.

[72] M. Xu and X. Li, "Subject property inference attack in collaborative learning," in *2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 1, 2020, pp. 227–231.

[73] M. Chase, E. Ghosh, and S. Mahloujifar, "Property inference from poisoning," 2021. [Online]. Available: https://arxiv.org/abs/2101.11073

[74] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Towards demystifying membership inference attacks," 2019. [Online]. Available: https://arxiv.org/abs/1807.09173

[75] Y. Mao, X. Zhu, W. Zheng, D. Yuan, and J. Ma, "A novel user membership leakage attack in collaborative deep learning," in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2019, pp. 1–6.

[76] J. Chen, J. Zhang, Y. Zhao, H. Han, K. Zhu, and B. Chen, "Beyond model-level membership privacy leakage: an adversarial approach in federated learning," in *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, 2020, pp. 1–9.

[77] J. Zhang, J. Zhang, J. Chen, and S. Yu, "Gan enhanced membership inference: A passive local attack in federated learning," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[78] W. Xia, Y. Li, L. Zhang, Z. Wu, and X. Yuan, "Cascade vertical federated learning towards straggler mitigation and label privacy over distributed labels," *IEEE Transactions on Big Data*, pp. 1–14, 2023.

[79] K. Fan, J. Hong, W. Li, X. Zhao, H. Li, and Y. Yang, "Flsg: A novel defense strategy against inference attacks in vertical federated learning," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 1816–1826, 2024.

[80] J. Sun, X. Yang, Y. Yao, and C. Wang, "Label leakage and protection from forward embedding in vertical federated learning," 2022. [Online]. Available: https://arxiv.org/abs/2203.01451

[81] H. Takahashi, J. Liu, and Y. Liu, "Eliminating label leakage in tree-based vertical federated learning," 2023. [Online]. Available: https://arxiv.org/abs/2307.10318

[82] O. Li, J. Sun, X. Yang, W. Gao, H. Zhang, J. Xie, V. Smith, and C. Wang, "Label leakage and protection in two-party split learning," 2022. [Online]. Available: https://arxiv.org/abs/2102.08504

[83] O. Goldreich, "Secure multi-party computation," *Manuscript. Preliminary version*, vol. 78, no. 110, pp. 1–108, 1998.

[84] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Advances in Cryptology — EUROCRYPT '99*, J. Stern, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 223–238.

[85] T. Elgamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE Transactions on Information Theory*, vol. 31, no. 4, pp. 469–472, 1985.

[86] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Commun. ACM*, vol. 21, no. 2, p. 120 – 126, feb 1978. [Online]. Available: https://doi.org/10.1145/359340.359342

[87] M. Bellare, V. T. Hoang, and P. Rogaway, "Foundations of garbled circuits," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, ser. CCS '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 784 – 796. [Online]. Available: https://doi.org/10.1145/2382196.2382279

[88] R. Chourasia, B. Enkhtaivan, K. Ito, J. Mori, I. T. ishi, and H. Tsuchida, "Knowledge cross-distillation for membership privacy," in *Proc. Priv. Enhancing Technol.*, vol. 2, 2022, p. 362 – 377.

[89] X. Tang, S. Mahloujifar, L. Song, V. Shejwalkar, M. Nasr, A. Houmansadr, and P. Mittal, "Mitigating membership inference attacks by Self-Distillation through a novel ensemble architecture," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 1433–1450.