Operations Research and Fuzziology 运筹与模糊学, 2022, 12(1), 1-15
Published Online February 2022 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2022.121001

Hans汉斯

机器学习成员推理攻击研究进展与挑战

高　婷

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2021年12月17日；录用日期：2022年1月19日；发布日期：2022年1月27日

摘　要

成员推理攻击通过对机器学习模型进行攻击可推断目标数据是否为训练数据集的成员，该攻击的日益完备给机器学习带来了严重的隐私威胁。本文从机器学习模型的攻防基础理论出发，分析成员推理攻击关键技术模型，厘清成员推理攻击模型与隐私泄露风险之间的关系，以期保证数据的隐私安全并促进机器学习应用领域的发展。首先，介绍了成员推理攻击的敌手模型、定义、分类以及攻击模型的生成机理。其次，分类总结和对比分析了成员推理攻击的攻击算法。然后，介绍了成员推理攻击在现实生活中的应用，并对成员推理攻击的防御技术进行了分类概括和对比分析。最后，通过对比分析已有的成员推理攻击方案及其防御技术方法，对机器学习成员推理攻击的发展趋势以及数据隐私保护的未来研究挑战进行展望。该工作为解决数据的隐私泄露问题提供一定的理论基础，对推动机器学习应用领域的发展有一定意义。

关键词

机器学习，成员推理攻击，防御技术，隐私泄露

Research Progress and Challenges of Membership Inference Attacks in Machine Learning

Ting Gao

College of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Dec. 17th, 2021; accepted: Jan. 19th, 2022; published: Jan. 27th, 2022

Abstract

Membership inference attacks can infer whether the target data is a member of a training dataset by attacking machine learning model, and the increasingly complete attack model poses a serious

文章引用: 高婷. 机器学习成员推理攻击研究进展与挑战[J]. 运筹与模糊学, 2022, 12(1): 1-15.
DOI: 10.12677/orf.2022.121001

privacy threat to machine learning. Starting from the basic theory of attack and defense of machine learning models, this paper analyzed the key technical models and clarified the relationship between attack models and privacy leakage risks for ensuring the security of data privacy and promoting the development of machine learning applications field. Firstly, this paper introduced the adversary model of membership inference attacks, definition, classification and generation mechanism of the model. Secondly, we summarized and analyzed various existing membership inference attack algorithms. Then, the application of membership inference attack in real life was introduced, and the defense techniques of membership inference attack was classified and compared. Finally, by comparing and analyzing the existing attack schemes and their defense technology methods, the development trend of membership inference attack in machine learning and the future research challenges of data privacy protection are prospected. This work provides a theoretical basis for solving the problem of data privacy leakage, which is of great significance for promoting the development of machine learning applications.

Keywords

Machine Learning, Membership Inference Attack, Defense Technology, Privacy Leakage

1. 引言

互联网的发展、硬件设备的更新、海量数据的收集以及智能算法的进步促使人工智能特别是机器学习[1]理论与技术快速发展，并广泛应用于诸多领域，如数据挖掘[2] [3]、计算机视觉[4] [5]、电子邮件过滤[6]、检测信用卡欺诈[7] [8] [9]和医学诊断[10] [11]等。通过收集大量相关数据进行分析，可以提高这些工作的效率。尽管机器学习技术及应用备受瞩目，产生了便利化、智能化的优势。但是，用户的生理特征、医疗记录、社交网络等大量个人敏感信息的收集使得蓬勃发展的机器学习技术的安全与隐私面临更加严峻的挑战。如2016年Yahoo据报道称被黑客盗取了至少5亿个用户账号信息；2017年微软Skype软件服务遭受DDOS攻击，导致用户无法通过平台进行通信；2020年《华盛顿邮报》报道称视频会议软件Zoom存在的重大安全漏洞。可见，机器学习应用领域中的数据隐私和安全问题已为社会的稳定带来严重的危害。

目前，机器学习安全与隐私的威胁主要分为4类：投毒攻击[12] [13] (poisoning attack)、对抗样本攻击[14] [15] (adversarial sample attack)、模型提取攻击[16] (model extraction attack)和模型逆向攻击[17] (model inversion attack)，如图1所示。其中，投毒攻击和模型逆向攻击发生在机器学习的训练阶段，前者主要在训练阶段投放恶意数据来降低模型的准确预测性能，后者主要通过逆向推理来获得训练集的信息，主要包括成员推理攻击和属性推理攻击；模型提取攻击和对抗样本攻击发生在机器学习的推理阶段，前者主要窃取模型的内部信息，后者主要在推理阶段通过添加干扰因子以生成对抗样本来欺骗模型。由图1可知，针对众多的攻击威胁，相应的防御措施也相继产生，比如，同态加密[18]、安全多方计算[19]和差分隐私[20]。这些防御技术能有效规避数据的安全与隐私威胁。

机器学习的训练有赖于训练数据集的数量与质量，而个人敏感数据的泄露正威胁着机器学习的广泛应用。模型逆向攻击可以从模型中恢复出训练数据，是目前机器学习所面临的主要隐私挑战，严重威胁机器学习隐私安全。该类攻击中包含的成员推理攻击能够成功推断某一目标样本是否是目标训练数据集的成员，从而造成隐私泄露。该攻击已在各种数据域中成功实施，如生物医学数据[21] [22] [23]，移动位置数

据[24]等。2008 年，Homer 等[21]通过分析目标人群和从公共来源获得的参考人群的信息汇总情况，成功推断出目标人群的疾病案例。随后，Backes 等[23]将此攻击扩展到其他类型的生物医学数据，如对将 microRNA 用于科学研究的个人进行成员推理。此外，聚合移动轨迹也被证明容易受到成员推理攻击[24]。

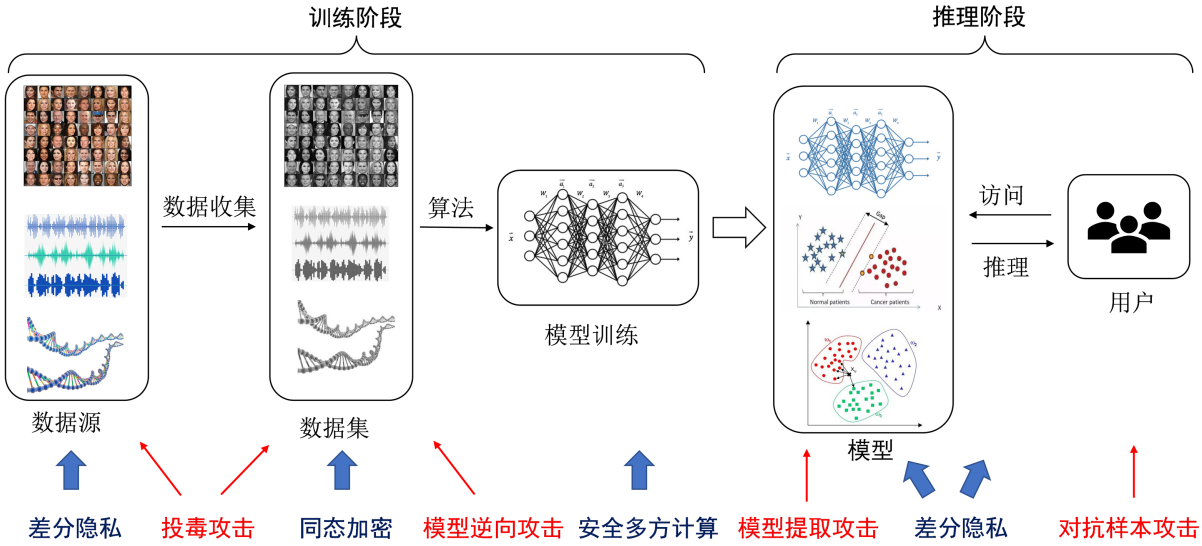

Figure 1. Security and privacy threats of machine learning
图 1. 机器学习的安全与隐私威胁

鉴于不同学者所处的研究领域不同，解决问题的角度不同，针对成员推理攻击与防御研究的侧重点不同。因此，本文从机器学习模型的攻防基础理论出发，分析关键技术模型，厘清成员推理攻击模型与隐私泄露风险之间的关系，对保证数据隐私的安全以及机器学习应用领域的发展有着重要的意义。本文第 2 节简述了成员推理攻击的敌手模型、定义、分类以及生成机理。第 3、4 节分别分析了不同类型的成员推理攻击算法的攻击手段和应用现状。第 5 节梳理和总括了不同攻击手段的防护策略及其成功的深层原因。最后，第 6、7 节总结全文，并对未来提出展望。

2. 成员推理攻击

本节重点针对成员推理攻击已有研究进行总结和归纳。

2.1. 敌手模型

在机器学习安全中，常常利用敌手模型来刻画一个敌手的强弱。Barreno 等[25]在 2010 年考虑了攻击者能力、攻击者目标的敌手模型。Biggio 等[26]在 2013 年更进一步提出了包含敌手目标、敌手知识、敌手能力和敌手策略的敌手模型。从这 4 个维度来刻画成员推理攻击敌手，能够比较系统地描述出敌手的威胁程度，如表 1 所示。

Table 1. Adversary model in membership inference attack
表 1. 成员推理攻击中的敌手模型

敌手模型	描述
敌手目标	可用性和隐私性的破坏
敌手知识	黑盒、白盒

Continued

敌手能力	强敌手：可以干预模型训练、访问训练数据集和收集中间结果等； 弱敌手：只能通过攻击手段获取模型信息或者训练数据信息。
敌手策略	训练阶段：模型逆向攻击 预测阶段：对抗攻击 + 成员推理攻击、模型提取攻击+成员推理攻击

2.2. 定义与模型

成员推理攻击通过分析目标模型系统来获取训练数据的成员关系信息，这是隐私泄露中最普遍的一类攻击。该攻击是一种判断给定目标数据是否用于训练目标模型的攻击方法，通过这一攻击方法，攻击者可以推测出模型训练集的信息，结构如图 2 所示。

由图 2 可知，原始数据集训练的目标模型在应用平台上运行，攻击者冒充用户去访问目标模型，获得一定的信息和敌手知识来构建攻击模型用于推理任意给定数据是否是目标模型的训练集成员。

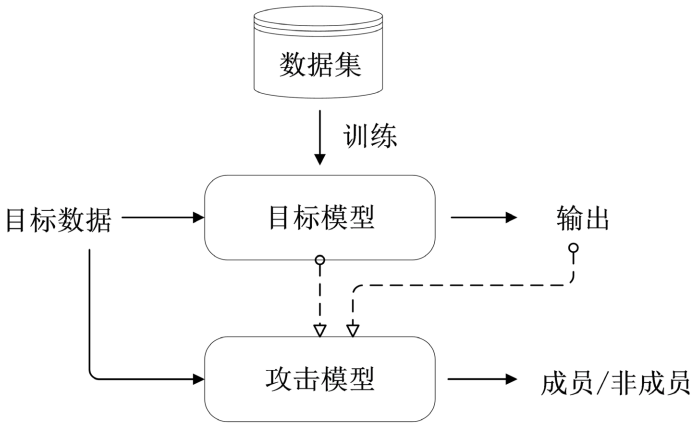

Figure 2. The model of membership inference attack
图 2. 成员推理攻击模型

2.3. 攻击模型的分类

针对目前成员推理攻击已有研究，根据不同标准分为以下几类，如表 2 所示。

Table 2. Types of membership inference attacks
表 2. 成员推理攻击的类别

敌手知识	攻击方法	攻击模式	目标模型	类型
黑盒	影子技术攻击	被动攻击	分类模型/深度学习/图神经网络/迁移学习	集中学习
	基线攻击	被动攻击	分类模型	集中学习
	标签攻击	被动攻击	分类模型/深度学习	集中学习
	转移攻击	被动攻击	分类模型	集中学习
白盒	白盒攻击	被动攻击/主动攻击	深度学习/生成对抗网络	集中学习/联邦学习

由表 2 可知，根据攻击者对目标模型信息的掌握程度即敌手知识，成员推理攻击可以分为两类：

黑盒攻击[27]-[33]和白盒攻击[34] [35]。黑盒攻击是指攻击者只能通过相应的 API 获得模型输出结果。即在黑盒攻击中对于输入 x，攻击者只能看到模型的输出 $f(x;W)$。但是无法获得模型的中间结果；白盒攻击是指攻击者可以访问目标模型结构、训练参数、模型内部输出结果，训练数据分布以及部分相关数据信息，借助所有获取信息来部署攻击模型。也有部分学者研究了敌手已知目标对象防御信息的水晶盒攻击[36]。

根据攻击者的攻击模式，分为强敌手和弱敌手，即主动攻击和被动攻击。前者能够干预目标模型训练、在联邦学习中充当参与方访问训练数据集，攻击者拥有改变目标模型训练过程中间数据的能力；后者只能观察训练过程的数据变化，通过被动获取模型接口输出信息进行攻击。

鉴于不同的攻击类型，成员推理攻击主要基于集中学习和联邦学习两类。集中学习是一类传统的训练模式，将数据集集中存储来训练目标模型。而在联邦学习模型下，每个参与方在当地存储个人数据并训练，它们之间通过中央参数服务器进行梯度交换，共同训练中央模型。在这种模式下，攻击者可能是中央参数服务器或当地某一参与方。

早期的成员推理攻击主要针对机器学习，后来随着图像、文本、以及知识图谱等数据的广泛应用，面向迁移学习、深度学习、图神经网络以及生成模型的成员推理攻击也应运而生，攻击范围更广泛，造成的隐私风险更高。

2.4. 攻击的生成机理

成员推理攻击能够成功推断主要在于目标模型存在过拟合的缺陷，目标模型能够记住训练数据的隐含特征，使得攻击者能够成功推断目标数据的成员关系。此外，异常数据的出现、数据的分布特征以及模型训练时的中间过程等使得攻击者能够侦破目标并成功实施攻击。

1) 过拟合[28] [31] [33]：成员推理攻击是指攻击者能够区分出目标模型的训练集与测试集，进而确定给定目标数据是否属于训练集成员。机器学习模型的过拟合问题使得该模型能够高精度地预测训练集但对测试集的预测能力较低。因此，因过拟合而导致的泛化性能差的模型易受到成员推理攻击。

2) 离群点[37] [38] [39]：当训练集数据中存在的关键离群点不具有代表性，即其分布与测试集数据的分布不同，此时训练集训练出来的模型并不能完美地适应测试数据集，从而导致模型的训练集与测试集易于被区别从而使得成员推理攻击能够成功。

3) 数据与模型的多因子影响[40] [41]：除了已提及的影响因素，数据的隐私泄露还与原始数据的特征以及攻击过程中的内在结构有关，比如：影子数据集大小、类平衡、特征平衡、模型配置(参数梯度、梯度范数)、数据不可区分性偏差以及涉及的数据熵、互信息量等。成员推理攻击并非仅由单个因素影响，而是多个因素共同作用。

3. 攻击算法

成员推理攻击已在各种数据域中成功实施，其中，机器学习成员推理攻击主要涉及黑盒知识和白盒知识两类，具体如下。

3.1. 黑盒知识

成员推理攻击的大部分工作[31] [32] [33] [34] [38] [39]都使用了黑盒模型。Shokri 等[31]首次提出了针对机器学习模型的成员推理攻击，通过该攻击成功判断特定病人是否出院。后来，Salem 等[32]通过逐渐放松 Shokri 等的假设提出另一种攻击，该攻击能够实现较好的准确率和召回率。此外，还有其他几种面向机器学习模型的基于置信度的成员推理攻击陆续出现，例如联合学习、生成对抗网络、自然语言处理、迁移学习和计算机视觉分割等。在基于决策的攻击领域[33] [38] [39]，Yeom 等[33]定量分析了训练

集和测试集的攻击性能与损失之间的关系，提出了第一个基于决策的攻击，即基线攻击。Choo 等[38]提出了一种类似于边界攻击的方法。

具体而言，黑盒攻击主要包含影子技术攻击、基线攻击、标签攻击以及转移攻击等方案。

1) 影子技术攻击

最初的成员推理攻击是 Shokri 提出的针对机器学习的成员推理攻击，即影子技术攻击[31]。该方案需要借助影子技术来模拟目标模型，并构建训练数据集来训练二分类攻击模型，从而实现成员推断，具体细节如图 3 所示。

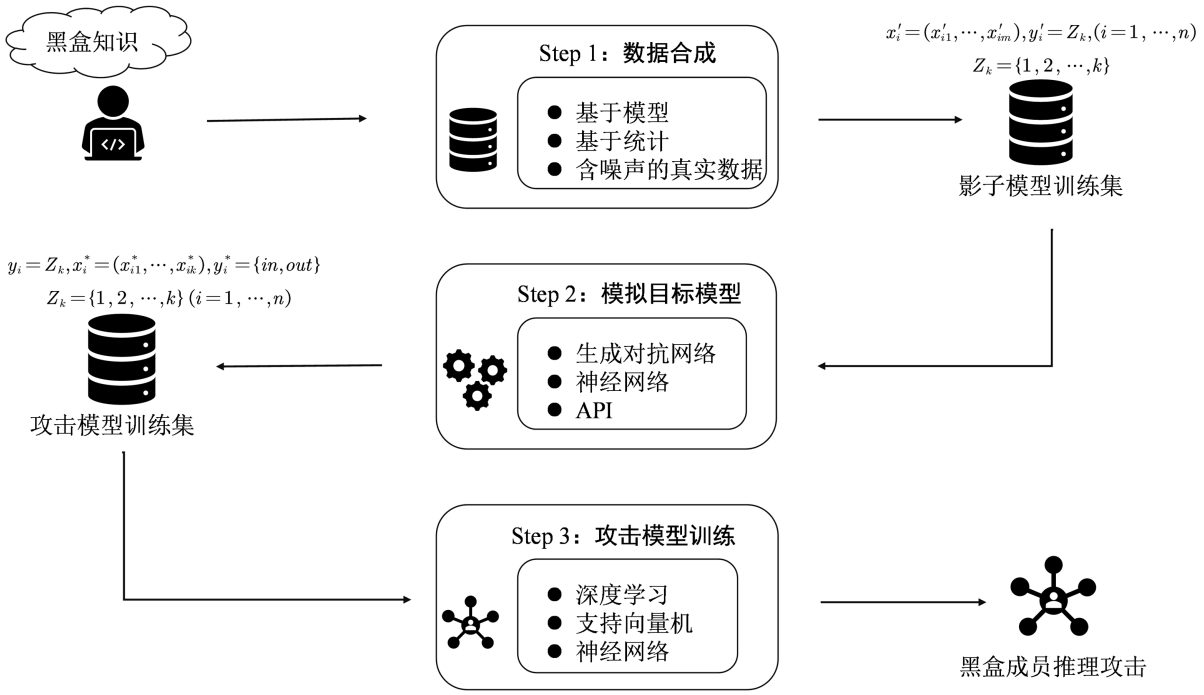

Figure 3. Black-box attack
图 3. 黑盒攻击

由图 3 可知，该方案主要分为三步：数据合成、影子模型模拟和攻击模型构建。

a) 数据合成：黑盒访问场景下，攻击者并不知道成员数据信息。故需要通过不同的统计算法去合成近似数据，比如基于模型、基于统计分布、基于噪声的真实数据。

b) 影子模型模拟：利用合成的相关数据去训练一个或者多个影子模型，该影子模型结构和目标模型结构一致，即在未知目标模型任何信息的情况下，影子技术模拟目标模型，通过分析模拟的影子模型来替代原始的目标模型。

c) 攻击模型构建：基于影子模型的数据集和目标模型的输出置信向量，并结合设定标签(如果数据 x 输于影子模型的训练集，则 $label=1$，否则 $label=-1$)训练一个二分类攻击模型，该模型能够判断给定目标数据点是否属于目标模型的训练数据集。

接着，Salem 等[32]放松了 Shokri 的假设条件，提出仅借助目标模型的输出结果进行阈值判别，就能实现较好的准确率和召回率，公式见(1)。该方法操作简便，效率不低，但仅适用于泛化性差的模型。

$$\mathcal{A}(x)=\begin{cases}1 & \text{if } \phi(x)>\tau\\ -1 & \text{otherwise}\end{cases} \tag{1}$$

借助影子技术的黑盒攻击起初是针对云平台上的机器学习模型 API 接口，接着延伸到深度学习、迁移学习和图神经网络。在针对训练社交网络和蛋白质结构这些数据的图神经网络的影子技术攻击中[34]，合成数据以及影子模型可以与目标系统不一致，甚至是针对泛化性能良好的模型，也能实现效果不错的攻击。这是因为在图神经网络中，实例之间的连通性增加了 GNN 模型对隐私攻击的脆弱性。

2) 基线攻击

Yeom 等[33]在 2018 年提出的基线攻击是通过数据样本是否被正确分类来进行成员推断。若目标数据被错误分类，则认定该数据为非成员数据，反之为成员数据，公式如下。

$$\mathcal{A}(x)=sign\left(f(x),y\right)=\begin{cases}-1 & \text{if } \arg\max\limits_{c} f(x)\neq y\\ 1 & \text{otherwise}\end{cases} \tag{2}$$

基线攻击的攻击强度随模型的过拟合情况呈正相关，针对存在较大的泛化差距的模型，攻击性能高，成本低，但是对泛化性良好的模型失效。

3) 标签攻击

Choo 等[38]提出了一种类似于边界攻击的方法，在黑盒设置下，只借助目标模型的输出标签来进行攻击，该攻击基于训练集样本相比测试集样本更难被扰动的原理下进行。

基于标签的成员推理攻击方案可划分为以下 3 个阶段。

a) 对抗样本生成：以目标模型的预测标签作为模型的输入，采用 FGSM、C&W、hopskipjump 等对抗样本技术对目标进行决策变动，生成对抗样本。

b) 扰动映射：计算对抗样本与原始目标之间的欧式距离，将扰动难度映射到距离范畴来寻找目标模型的训练数据和测试数据的预测差异。

c) 成员推断：将预测差异进行逻辑判别获得细粒度的成员信号，以实现目标人群的成员推断，公式如下。

$$\mathcal{A}(x)=sign\left(d\left(x_{adv},x\right),\tau\right)=\begin{cases}1 & \text{if } d\left(x_{adv},x\right)\geq\tau\\ -1 & \text{otherwise}\end{cases} \tag{3}$$

其中，$\mathcal{A}(x)$为 1 时，代表 x 为目标成员，反之为非成员。

4) 转移攻击

文献[39]提出一类攻击，给定数据点(x,y)和从目标模型获取的置信向量 $f(x)$，计算交叉熵损失 $loss(x,y)=-\log\left(f(x)_y\right)$。

$$\mathcal{A}(x)=\begin{cases}1 & \text{if } loss(x,y)<\bar{z}_{loss}\\ -1 & \text{otherwise}\end{cases} \tag{4}$$

其中，$\bar{z}_{loss}$ 为训练样本集的平均损失值，通过影子训练数据集获得。 $\mathcal{A}(x)$为 1 时，代表 x 为目标成员，反之为非成员。

3.2. 白盒知识

在黑盒知识的攻击工作中，攻击者只能根据模型的输出来对训练数据进行攻击。然而训练过程的中间计算数据也包含了大量有关训练数据的信息。因此，前人在针对 GAN 的攻击工作中首次提出了白盒攻击，该攻击仅使用 GAN 鉴别器部分的输出，而无需鉴别器或生成器的学习权重即可完成攻击。此外，Nasr 等提出将成员推理攻击拓展到基于先验知识的白盒设置[35]，将从目标模型那获得的激活函数和梯度信息作为推断的特征，来进行成员推断，具体细节如图 4 所示。

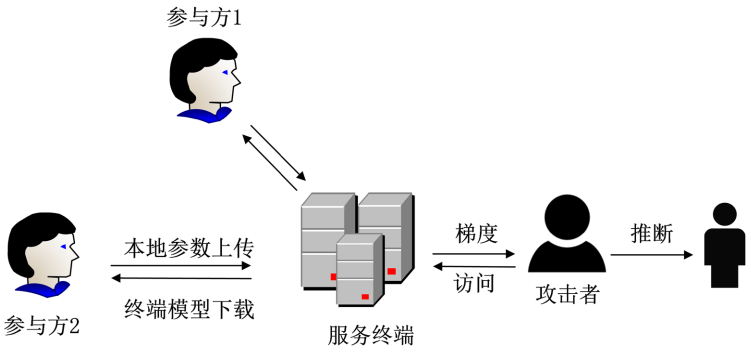

Figure 4. White-box attack
图 4. 白盒攻击

由图 4 可知，该方案基于目标模型根据训练集的数据进行微调更新使训练数据的损失梯度趋于零的原理，区分训练集数据和非训练集数据的梯度来实现成员推理。

对于一个目标模型 f，输入一个数据 x。攻击者在目标模型的前向传播计算中计算各个层的输出 $h_i(x)$，模型输出 $f(x)$ 和损失 $L(f(x;W),y)$。进一步可以通过反向传播计算各个层的梯度 $\frac{\partial L}{\partial W_i}$。然后，将计算得到的参数与 y 的 one-hot 向量一同组成了攻击模型的输入特征参数。

将输入特征分别输送到相应的 CNN 或者 FCN 进行特征提取，并将输出进行打包，然后再传给全连接层 FCN，最后得到推理攻击的结果。其中，攻击模型由卷积神经网络(CNN)和全连接网络(FCN)两部分组成。

另外，Long 等[37]提出了一种针对泛化良好模型的成员推理攻击 GMIA。在该攻击中，并非所有的数据都易收到成员攻击，攻击者需要寻找易受攻击的异常数据点进行成员与非成员的区分，进而完成攻击。

3.3. 算法对比

本部分对比了上述提到的部分算法，对目前已有的成员推理攻击算法进行了细致的归纳，具体情况如表 3 所示。

Table 3. Comparison of membership inference attack algorithms
表 3. 成员推理攻击算法对比

	敌手知识	目标模型	类型	假设条件			攻击精度	
				影子模型	数据分布	模型结构	数据集	准确性(%)
[31]	黑盒	分类模型	独立模型	是	是	是	CIFAR100	92.8
[32]	黑盒	CNN	独立模型	否	否	否	CIFAR100	85.7
[34]	白盒	分类模型	联合学习	/	/	/	Yelp-health	75.0
[35]	白盒	分类模型	联邦学习	/	/	/	CIFAR100	85.1
[39]	黑盒	DL	独立模型	否	是	否	CIFAR10	88.0
[42]	黑/白盒	生成模型	独立模型	否	否/是	否	LFW	61.0/94.3
[43]	黑盒	NN	独立模型	是	否	否	Tweet(4)	64.8

4. 成员推理攻击的应用现状

鉴于成员推理攻击可以推理出模型训练集中是否含有某一数据，其可以用于测试用户的数据有没有被未授权使用，以及用于疾病监测和安全监管。此外，在尚未遭受攻击时对机器学习系统进行风险评估和隐私加固，提高系统抵御潜在攻击的能力。

4.1. 审计判别

Miao 等[44]设计了一个语音审计模型，其能够检测用户的语音数据是否在目标模型的训练集中的，从而推断出用户数据有没有被未授权使用。该模型重点关注的是用户层面的成员推理，推理某个用户的数据是否在非自愿参与训练的情况下被目标模型非法使用，从而维护用户权益，使其能够对目标系统模型进行审计监管。同理，Song 等[45]提出了一种用于文本生成模型的审计模型，使用成员推理来判断用户数据有没有被未授权使用。

4.2. 疾病预测

成员推理攻击应用于医疗数据用于疾病监测[21] [22] [23] [36]，比如 Homer 等[21]通过将目标个人的概况和案例研究的汇总以及从公共来源获得的参考人群的汇总进行比较，可以了解目标个人是否属于与某种疾病相关的案例研究组中。此外，对于由艾滋病患者数据构建的诊断模型，若某人的医疗数据被推断是该模型的训练数据，便意味着此人可能患有艾滋病。

4.3. 安全监管与知识产权

成员推理攻击还可应用于用户的信用监测[47] (外卖平台一户多用)；聚合位置监测[24]；评估系统(平台)发布数据之前的隐私保护质量；监管部门监测用户信息是否被非法滥用，便于用户维权。甚至，成员推断还损害了模型提供者对培训数据集的知识产权。

5. 防御策略

为抵御多样的成员推理攻击，其针对性的防御方案同样引发了研究者们的高度关注和重点研究。

5.1. 防御技术

成员推理攻击已经威胁到了训练集数据的成员隐私问题，针对成员推理的防御可以分为三类：1) 基于正则化的防御[31] [32] [48] [49] [50]：直接采用正则化技术来构建防御，如 L2 正则化、dropout、模型堆叠和 min-max；2) 基于对抗性攻击的防御：通过对抗性攻击来保护受害者模型；3) 基于差分隐私防御[51]：在训练数据输入、目标函数、模型梯度以及输出过程中添加扰动噪声，来降低成员的隐私泄漏。

下面重点介绍一些最新的防御技术及其优缺点。

5.1.1. min-max game

Nasr 引入了一种博弈思想来训练具有成员资格私有性的模型[48]，从而确保模型在其训练数据与其他数据点的预测之间没有可区分性。该隐私机制针对最强的推理攻击，将隐私损失降到最低，将分类损失降至最低。该机制优化了最小－最大目标函数，公式如下。

$$\min_{f}\left(L_D(f)+\lambda\max_{h}G_{f,D,D'}(h)\right) \tag{5}$$

其中，$L_D(f)=loss\left(f(x),y\right)$，$G_{f,D,D'}(h)=\frac{1}{2}\log\left(h\left(x,y,f(x)\right)\right)+\frac{1}{2}\log\left(1-h\left(x',y',f(x')\right)\right)$，$f$ 为分类器，h 为攻击模型。该算法的隐私保护机制不仅保护成员隐私，而且可以显著防止过度拟合。

5.1.2. mem-guard

这是第一个针对成员推断提供正式效用损失保证的防御[49]。其基本思想是在 ML 模型的置信度分数中添加精心设计的噪声，以误导成员分类器。即把噪声向量 $\boldsymbol{n}$ 添加到置信得分向量 $\boldsymbol{s}$ 的概率中，对成员推理攻击提供效用－损失保证的防御[41]。该算法需要找到满足唯一效用－损失约束的噪声向量，公式如下。

$$\begin{aligned}&\mathcal{M}^{*}=\arg\min_{\mathcal{M}}\left|E_{\mathcal{M}}\left(g\left(\boldsymbol{s}+\boldsymbol{n}\right)\right)-0.5\right|\\&\text{subject to}:\arg\max_{j}\left\{s_j+n_j\right\}=\arg\max_{j}\left\{s_j\right\}\\&\qquad E_{\mathcal{M}}\left(g\left(\boldsymbol{s}+\boldsymbol{n}\right)\right)\le\epsilon\\&\qquad s_j+n_j\ge 0,\forall j\\&\qquad\sum_{j}s_j+n_j=1\end{aligned}\tag{6}$$

其中，g 为防御分类器，ϵ 为置信得分向量变化预算，$\mathcal{M}$ 为噪声添加机制函数。该算法是针对黑盒攻击的防御方法，通过对目标模型得到的置信得分向量以一定的概率添加噪声得到一个随机噪声添加机制，并且让防御者模拟攻击者的攻击分类器形成防御分类器，进而提出优化问题并且求解。实验证明，mem-guard 强于 min-max game 和模型堆叠。

5.1.3. 差分隐私

Chen [51]提出的差分隐私防御技术是通过对模型的权值进行扰动，进而有效地保护模型的隐私。较小的隐私预算提供了更强的隐私保障，但损失了更多的模型准确性，表明差分隐私的成员隐私和模型效用之间的平衡难以解决。实验将隐私预算和目标模型准确性之间的关系建模为一个对数型曲线，在拐点附近找到隐私和效用之间的权衡预算。此外，结合模型稀疏性的差分隐私能够显著降低成员推理攻击的风险。

5.1.4. 其他防御技术

Li [52]提出的 MMD + Mix-up 算法通过使用训练集和验证集的 softmax 输出经验分布之间的最大平均差异作为正则化器添加到模型的损失函数中，使得成员与非成员样本之间的分布尽可能小，从而防御攻击。

Wang [53]提出的 MIA-pruning defense 是一种剪枝算法。该剪枝算法通过找到一个能够防止 MIA 泄露隐私的子网络，并达到与原始 DNN 相当的精度。该算法能够在降低模型存储和计算复杂度的同时，降低模型的精度损失，且优于差分隐私、min-max game 技术。其他的 PAR-GAN defense [54]是通过降低模型的泛化误差来进行模型防御。

综上，表 4 对比和总结了现有一些典型的成员推理攻击防御方案。

Table 4. Defense technology
表 4. 防御技术

方法	技术特点	优点	缺点
数据降维[27]	缩小训练集中核心特征的动态范围	隐私性高	计算开销大
随机失活[32]	一定比例随机失活神经元来泛化模型	隐私性高	治标不治本
模型堆叠[32]	多模型训练，降低模型记忆能力，避免过拟合	隐私性高	治标不治本

Continued

min-max game [48]	最大化攻击成功率的基础上，最小化模型的预测损失	隐私性和可用性好	数据先验知识要求，强大的计算开销
mem-guard [49]	预测置信度向量上添加噪声得到对抗样本	隐私性和效率较高	使用环境有限
知识蒸馏[50]	利用迁移知识使得成员和非成员的损失相近	隐私性高	可用性不高
差分隐私[51]	噪声扰动目标函数	隐私性和效率较高	影响分类器的性能，可用性不高

5.2. 防御成功的原因

面向目前已有的攻击，提出了一系列基于正则化、对抗样本以及差分隐私的防御手段，其防御成功的深层原因如下。

1) 基于正则化的防御：针对模型的过拟合问题所造成的高精度成员推理攻击，基于正则化的防御策略可以通过提升测试集的精度来降低模型的过拟合，提升模型的泛化能力，保证机器学习算法的稳健性，进而防御成员推理攻击。但是，某些情况下，模型系统上部署的这类防御技术并不能完全消除泛化差距。

2) 基于对抗性攻击的防御：鉴于基于预测置信度的成员推理攻击是通过预测置信度来完成攻击，该防御策略是通过向输出的置信度向量添加噪声扰乱成员与非成员之间的区别界限，进而达到相应的防御效果。

3) 基于差分隐私防御：该策略是在训练数据输入、目标函数、模型梯度以及输出过程中添加扰动噪声，降低目标模型对训练数据的记忆从而降低成员泄漏的风险。

6. 挑战与建议

随着人工智能研究和应用的逐步深入，机器学习算法的特殊性给用户数据和网络模型的隐私保护带来巨大挑战，迫切需要进一步考虑更高的安全及隐私威胁，并提出更强适应范围更广的防御手段，提升机器学习模型的利用率。接下来，分析成员推理攻击以及防御的研究挑战，并就其未来的研究挑战进行展望。

1) 研究基于白盒知识的高效机器学习成员推理攻击

目前，基于黑盒知识的成员推理攻击在大多数数据上都能实现不错的推理性能，且广泛应用于不同研究领域，但效率低于白盒攻击且存在一定的限制条件。例如，基于黑盒知识的影子技术的攻击性能受模型的泛化性影响且受限于数据分布以及模型结构类似假设要求；低成本的应用场景给标签攻击带来了更多的发展机会，但仅限于深度学习，其他应用领域并未涉及；转移攻击和借助阈值的黑盒攻击是一种简单且高效的攻击方法，但是阈值的确定需要借助影子技术来获取，限制条件过多不利于实际场景展开。因此，研究基于白盒知识的高效成员推理攻击是一个亟待解决的问题。

2) 设计适用于机器学习各种算法的通用成员推理攻击机制

一方面，基于黑盒知识的攻击大多基于过拟合导致效率不高且稳定性差，必须深入研究数据泄露的深层机理。另一方面，针对泛化性良好模型的白盒攻击在实际应用中的涉及面有限且攻击者在白盒知识下对联邦学习实施攻击时并不能有效推理出具体哪一方的数据信息参与。因此，需要结合有效的属性推断设计适用于机器学习各种算法的通用成员推理攻击机制。

3) 提出针对非欧式空间数据的更加切实可行的攻击方案

现有的成员推理攻击大多数集中于基于欧氏空间数据(比如图像和文本)训练的机器学习模型上。但真实世界的数据大多以图表的形式出现，比如社交网络和蛋白质结构。最新研究表明，图神经网络可用于

处理这类数据，但是针对这方面的成员攻击仅限于影子技术手段。基于这些数据的机器学习模型上的隐私攻击未得到充分研究。因此，在不影响在线社交网络用户体验的情况下针对非欧式空间的数据隐私是一个很有前途的研究方向。

4) 实现隐私性、高效性和可用性之间的最佳平衡

机器学习中训练数据的隐私性、模型的高效性和可用性之间相互矛盾。例如，基于差分隐私的防御方法更加私密和高效，但由于添加了噪声扰动仅能实现次优的效用-隐私平衡；基于安全多方计算的防御方法在隐私性和可用性方面都很高，但由于双方的相互作用和高通信开销导致效率低下。因此，有必要建立隐私保护机制的多维评价体系，模拟不同模式和不同攻击方法下的三者关系，实现不同场景下三者之间的权衡优化。

5) 建立统一的隐私泄露度量标准

机器学习成员推理攻击研究中，如何度量机器学习模型的隐私泄露风险，是评估攻击性能体系中的一个重要问题。目前，一些学者对隐私量化研究工作已经展开，但还是比较零散，更多的是针对某个单一特定领域，其适用范围有限。加上隐私泄露涉及诸多因素，尚未形成统一的模式和制度。因此，建立统一的隐私泄露衡量标准和完善的隐私披露风险分析与评价机制，是机器学习中需要进一步研究的课题。

6) 优化传统的数据隐私保护方案

基于正则化、差分隐私和游戏博弈的方法可以降低成员推理攻击的隐私泄露。但是，考虑隐私数据的敏感性、模型强大的记忆能力，以及对训练数据完整性和机密性的要求，可以结合密码学、匿名、对抗性正则化和差分隐私等混合方法对传统的隐私保护防御方法进一步优化。

7. 总结

本文首先介绍机器学习面临的安全与隐私威胁现状，并分析数据隐私威胁中的成员推理攻击问题。其次，比较分析现有的成员推理攻击方式，以及其应用现状。然后，对比分析成员推理攻击的常用隐私保护方法及其防御成功的深层机理。最后，通过比较分析现有数据隐私保护方法的不足，讨论针对成员推理攻击隐私保护研究的挑战，为今后更强大的攻击做准备。

参考文献

[1] Jordan, M.I. and Mitchell, T.M. (2015) Machine Learning: Trends, Perspectives, and Prospects. *Science*, **349**, 255-260. https://doi.org/10.1126/science.aaa8415

[2] 廖国辉, 刘嘉勇. 基于数据挖掘和机器学习的恶意代码检测方法[J]. 信息安全研究, 2016, 2(1): 74-79.

[3] 韩莹, 李姗姗, 陈福明. 基于机器学习的地震异常数据挖掘模型[J]. 计算机仿真, 2014, 31(11): 319-322.

[4] Chen, X., Xiang, S., Liu, C.L., *et al.* (2014) Vehicle Detection in Satellite Images by Hybrid Deep Convolutional Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, **11**, 1797-1801. https://doi.org/10.1109/LGRS.2014.2309695

[5] Chen, S., Wang, H., Xu, F., *et al.* (2016) Target Classification Using the Deep Convolutional Networks for SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, **54**, 4806-4817. https://doi.org/10.1109/TGRS.2016.2551720

[6] Launchbury, J., Archer, D., DuBuisson, T., *et al.* (2014) Application-Scale Secure Multiparty Computation. In: Shao, Z., Ed., *European Symposium on Programming Languages and Systems*, Springer, Berlin, Heidelberg, 8-26. https://doi.org/10.1007/978-3-642-54833-8_2

[7] 凌晨添. 进化神经网络在信用卡欺诈检测中的应用[J]. 微电子学与计算机, 2011, 28(10): 14-17.

[8] Fu, K., Cheng, D., Tu, Y., *et al.* (2016) Credit Card Fraud Detection Using Convolutional Neural Networks. In: Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M. and Liu, D., Eds., *International Conference on Neural Information Processing*, Springer, Cham, 483-490. https://doi.org/10.1007/978-3-319-46675-0_53

[9] Roy, A., Sun, J., Mahoney, R., *et al.* (2018) Deep Learning Detecting Fraud in Credit Card Transactions. 2018 *Systems*

and Information Engineering Design Symposium (*SIEDS*), Charlottesville, VA, 27 April 2018, 129-134. https://doi.org/10.1109/SIEDS.2018.8374722

[10] Acharya, U.R., Oh, S.L., Hagiwara, Y., *et al.* (2018) Deep Convolutional Neural Network for the Automated Detection and Diagnosis of Seizure Using EEG Signals. *Computers in Biology and Medicine*, **100**, 270-278. https://doi.org/10.1016/j.compbiomed.2017.09.017

[11] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., *et al.* (2017) Computer Aided Decision Making for Heart Disease Detection Using Hybrid Neural Network-Genetic Algorithm. *Computer Methods and Programs in Biomedicine*, **141**, 19-26. https://doi.org/10.1016/j.cmpb.2017.01.004

[12] Jagielski, M., Oprea, A., Biggio, B., *et al.* (2018) Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. 2018 *IEEE Symposium on Security and Privacy* (*SP*), San Francisco, CA, 20-24 May 2018, 19-35. https://doi.org/10.1109/SP.2018.00057

[13] Liu, Y., Ma, S., Aafer, Y., *et al.* (2018) Trojaning Attack on Neural Networks. *Proceedings of the* 25*th Annual Network and Distributed System Security Symposium*, San Diego, CA, 18-21 February 2018, 214-229. https://doi.org/10.14722/ndss.2018.23291

[14] Szegedy, C., Zaremba, W., Sutskever, I., *et al.* (2013) Intriguing Properties of Neural Networks. arXiv:1312.6199

[15] Papernot, N., McDaniel, P., Jha, S., *et al.* (2016) The Limitations of Deep Learning in Adversarial Settings. 2016 *IEEE European Symposium on Security and Privacy* (*EuroS&P*), Saarbruecken, 21-24 March 2016, 372-387. https://doi.org/10.1109/EuroSP.2016.36

[16] Tramèr, F., Zhang, F., Juels, A., *et al.* (2016) Stealing Machine Learning Models via Prediction APIs. *Proceedings of the 25th USENIX Conference on Security Symposium*, Austin, TX, 10-12 August 2016, 601-618.

[17] Fredrikson, M., Jha, S. and Ristenpart, T. (2015) Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. *Proceedings of the* 22*nd ACM SIGSAC Conference on Computer and Communications Security*, Denver, CO, 12-16 October 2015, 1322-1333. https://doi.org/10.1145/2810103.2813677

[18] Gentry, C. (2009) Fully Homomorphic Encryption Using Ideal Lattices. *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, Bethesda, MD, 31 May 2009-2 June 2009, 169-178. https://doi.org/10.1145/1536414.1536440

[19] Jagannathan, G. and Wright, R.N. (2005) Privacy-Preserving Distributed k-Means Clustering over Arbitrarily Partitioned Data. *Proceedings of the* 11*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Chicago, IL, 21-24 August 2005, 593-599. https://doi.org/10.1145/1081870.1081942

[20] Jayaraman, B. and Evans, D. (2019) Evaluating Differentially Private Machine Learning in Practice. *Proceedings of the 28th USENIX Conference on Security Symposium*, Santa Clara, CA, 14-16 August 2019, 1895-1912.

[21] Homer, N., Szelinger, S., Redman, M., *et al.* (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genetics*, **4**, e1000167. https://doi.org/10.1371/journal.pgen.1000167

[22] Hagestedt, I., Zhang, Y., Humbert, M., *et al.* (2019) MBeacon: Privacy-Preserving Beacons for DNA Methylation Data. *Proceedings of the* 26*th Annual Network and Distributed System Security Symposium*, San Diego, CA, 24-27 February 2019, 72-87. https://doi.org/10.14722/ndss.2019.23064

[23] Backes, M., Berrang, P., Humbert, M., *et al.* (2016) Membership Privacy in MicroRNA-Based Studies. *Proceedings of the* 2016 *ACM SIGSAC Conference on Computer and Communications Security*, Vienna, 24-28 October 2016, 319-330. https://doi.org/10.1145/2976749.2978355

[24] Pyrgelis, A., Troncoso, C. and De Cristofaro, E. (2018) Knock Knock, Who's There? Membership Inference on Aggregate Location Data. *Proceedings of the* 25*th Network and Distributed Systems Security Symposium*, San Diego, CA, 18-21 February 2018, 199-213. https://doi.org/10.14722/ndss.2018.23183

[25] Barreno, M., Nelson, B., Joseph, A.D., *et al.* (2010) The Security of Machine Learning. *Machine Learning*, **81**, 121-148. https://doi.org/10.1007/s10994-010-5188-5

[26] Biggio, B., Fumera, G. and Roli, F. (2013) Security Evaluation of Pattern Classifiers under Attack. *IEEE Transactions on Knowledge and Data Engineering*, **26**, 984-996. https://doi.org/10.1109/TKDE.2013.57

[27] Hui, B., Yang, Y., Yuan, H., *et al.* (2021) Practical Blind Membership Inference Attack via Differential Comparisons. arXiv:2101.01341. https://doi.org/10.14722/ndss.2021.24293

[28] Li, J., Li, N. and Ribeiro, B. (2020) Membership Inference Attacks and Defenses in Supervised Learning via Generalization Gap. arXiv:2002.12062

[29] Song, L., Shokri, R. and Mittal, P. (2019) Privacy Risks of Securing Machine Learning Models against Adversarial Examples. *Proceedings of the* 2019 *ACM SIGSAC Conference on Computer and Communications Security*, London,

11-15 November 2019, 241-257. https://doi.org/10.1145/3319535.3354211

[30] Yang, Z., Shao, B., Xuan, B., *et al.* (2020) Defending Model Inversion and Membership Inference Attacks via Prediction Purification. arXiv:2005.03915

[31] Shokri, R., Stronati, M., Song, C., *et al.* (2017) Membership Inference Attacks against Machine Learning Models. 2017 *IEEE Symposium on Security and Privacy*, San Jose, CA, 22-26 May 2017, 3-18. https://doi.org/10.1109/SP.2017.41

[32] Salem, A., Zhang, Y., Humbert, M., *et al.* (2019) ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. *Annual Network and Distributed System Security Symposium*, San Diego, CA, 24-27 February 2019, 243-260. https://doi.org/10.14722/ndss.2019.23119

[33] Yeom, S., Giacomelli, I., Fredrikson, M., *et al.* (2018) Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. 2018 *IEEE* 31*st Computer Security Foundations Symposium*, Oxford, 9-12 July 2018, 268-282. https://doi.org/10.1109/CSF.2018.00027

[34] Melis, L., Song, C., De Cristofaro, E., *et al.* (2019) Exploiting Unintended Feature Leakage in Collaborative Learning. 2019 *IEEE Symposium on Security and Privacy*, San Francisco, CA, 19-23 May 2019, 691-706. https://doi.org/10.1109/SP.2019.00029

[35] Nasr, M., Shokri, R. and Houmansadr, A. (2019) Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-Box Inference Attacks against Centralized and Federated Learning. 2019 *IEEE Symposium on Security and Privacy*, San Francisco, CA, 19-23 May 2019, 739-753. https://doi.org/10.1109/SP.2019.00065

[36] Yin, Y., Chen, K., Shou, L. and Chen, G. (2021) Defending Privacy Against More Knowledgeable Membership Inference Attackers. *Proceedings of the* 27*th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Singapore, 14-18 August 2021, 2026-2036. https://doi.org/10.1145/3447548.3467444

[37] Long, Y., Bindschaedler, V., Wang, L., *et al.* (2018) Understanding Membership Inferences on Well-Generalized Learning Models. arXiv:1802.04889

[38] Choo, C.A.C., Tramer, F., Carlini, N., *et al.* (2020) Label-Only Membership Inference Attacks. arXiv:2007.14321

[39] Li, Z. and Zhang, Y. (2021) Membership Leakage in Label-Only Exposures. *Proceedings of the* 2019 *ACM SIGSAC Conference on Computer and Communications Security*, Korea, 15-19 November 2021, 880-895. https://doi.org/10.1145/3460120.3484575

[40] Wang, C., Liu, G., Huang, H., *et al.* (2019) MIASec: Enabling Data Indistinguishability against Membership Inference Attacks in MLaaS. *IEEE Transactions on Sustainable Computing*, **5**, 365-376. https://doi.org/10.1109/TSUSC.2019.2930526

[41] Tonni, S.M., Vatsalan, D., Farokhi, F., *et al.* (2020) Data and Model Dependencies of Membership Inference Attack. arXiv:2002.06856

[42] Hayes, J., Melis, L., Danezis, G. and De Cristofaro, E. (2019) LOGAN: Membership Inference Attacks against Generative Models. *Proceedings on Privacy Enhancing Technologies*, **2019**, 133-152. https://doi.org/10.2478/popets-2019-0008

[43] Liu, G., Wang, C., Peng, K., *et al.* (2019) SocInf: Membership Inference Attacks on Social Media Health Data with Machine Learning. *IEEE Transactions on Computational Social Systems*, **6**, 907-921. https://doi.org/10.1109/TCSS.2019.2916086

[44] Miao, Y., Zhao, B.Z.H., Xue, M., *et al.* (2019) The Audio Auditor: Participant-Level Membership Inference in Voice-Based IoT. *CCS Workshop of Privacy Preserving Machine Learning*.

[45] Song, C. and Shmatikov, V. (2019) Auditing Data Provenance in Text-Generation Models. *Proceedings of the* 25*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, 4-8 August 2019, 196-206. https://doi.org/10.1145/3292500.3330885

[46] Fredrikson, M., Lantz, E., Jha, S., *et al.* (2014) Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. *Proceedings of the* 23*rd USENIX conference on Security Symposium*, San Diego, CA, 20-22 August 2014, 17-32.

[47] Danhier, P., Massart, C. and Standaert, F.X. (2020) Fidelity Leakages: Applying Membership Inference Attacks to Preference Data. *IEEE INFOCOM* 2020-*IEEE Conference on Computer Communications Workshops* (*INFOCOM WKSHPS*), Toronto, 6-9 July 2020, 728-733. https://doi.org/10.1109/INFOCOMWKSHPS50562.2020.9163032

[48] Nasr, M., Shokri, R. and Houmansadr, A. (2018) Machine Learning with Membership Privacy Using Adversarial Regularization. *Proceedings of the* 2018 *ACM SIGSAC Conference on Computer and Communications Security*, Toronto, 15-19 October 2018, 634-646. https://doi.org/10.1145/3243734.3243855

[49] Jia, J., Salem, A., Backes, M., *et al.* (2019) MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. *Proceedings of the* 2019 *ACM SIGSAC Conference on Computer and Communications Security*, London, 11-15 November 2019, 259-274. https://doi.org/10.1145/3319535.3363201

[50] Zheng, J., Cao, Y. and Wang, H. (2021) Resisting Membership Inference Attacks through Knowledge Distillation. *Neurocomputing*, **452**, 114-126. https://doi.org/10.1016/j.neucom.2021.04.082

[51] Chen, J., Wang, W.H. and Shi, X. (2020) Differential Privacy Protection Against Membership Inference Attack on Machine Learning for Genomic Data. *BIOCOMPUTING* 2021: *Proceedings of the Pacific Symposium*, Kohala Coast, 3-7 January 2021, 26-37. https://doi.org/10.1142/9789811232701_0003

[52] Li, J., Li, N. and Ribeiro, B. (2021) Membership Inference Attacks and Defenses in Classification Models. *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, USA, 26-28 April 2021, 5-16. https://doi.org/10.1145/3422337.3447836

[53] Wang, Y., Wang, C., Wang, Z., *et al*. (2021) Against Membership Inference Attack: Pruning is All You Need. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* (*IJCAI*-21), 3141-3147.

[54] Chen, J., Wang, W.H., Gao, H., *et al*. (2021) PAR-GAN: Improving the Generalization of Generative Adversarial Networks against Membership Inference Attacks. *Proceedings of the* 27*th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Singapore, 14-18 August 2021, 127-137. https://doi.org/10.1145/3447548.3467445