理解脑电图模型中的成员推断攻击

Mehrdad Khodapanahandeh[1], Selin Unal[2],Berkay Kemal Balioglu[2],Alireza Khodaie[2],Beren Semiz[3,4],M. Emre Gursoy[2,5]

[1]生物医学科学与工程系，工程学院，*Koc*。大学，伊斯坦布尔，*T¨urkiye*[2]计算机工程系，工程学院，*Koc*。大学，伊斯坦布尔，*T¨urkiye*[3]电气与电子工程系，工程学院，*Koc*。大学，伊斯坦布尔，*T¨urkiye*[4]*Koc*。大学转化医学研究中心（*KUTTAM*），*Koc*。大学，伊斯坦布尔，*T¨urkiye*[5]*Koc*。大学与 *Is*。*Bank*人工智能中心（*KUIS AI*），*Koc*。大学，伊斯坦布尔，*T¨urkiye*{*mkhodapanahandeh21*、*sunal22*、*bbalioglu23*、*akhodaie22*、*besemiz*、*emregursoy*}*@ku.edu.tr*

arXiv:2611.28354v1 [cs.CR] 20 Nov 2026

关键词: 成员推断攻击（MIA）、脑电图（EEG）、深度学习、隐私、泛化差距、卷积神经网络。

摘要: 脑电图（EEG）信号包含敏感的受试者特异性信息，因此基于这些信号训练的模型可能成为成员推断攻击（MIAs）的目标。尽管成员推断攻击已在其他领域得到广泛研究，但其在脑电图模型中的应用相对较少，且导致脑电图模型易受成员推断攻击影响的因素尚未得到充分理解。本文使用三个数据集、七种模型架构、四种攻击方法和两个指标，对脑电图模型上的成员推断攻击开展了系统的实证分析。结果表明，较差的模型泛化能力（较大的训练集与测试集损失差距和准确率差距）是成员推断攻击脆弱性最强的指标。更长的训练时间和更大的模型也会增加脆弱性。此外，在不同数据集之间，没有任何特定的头皮区域、频带、通道间相关模式，或脑电图数据的时域与频谱表示能够持续解释成员推断攻击脆弱性。最后，受试者识别和留一受试者交叉实验表明，更依赖受试者特定表示的模型更容易受到成员推断攻击。总体而言，我们的研究结果确定了脑电图模型中成员信息泄露的关键驱动因素，并可为开发抗成员推断攻击脑电系统提供指导。

1 引言

脑电图已成为最广泛使用的无创

脑活动监测方式之一。它支撑着从癫痫和神经系统疾病的临床诊断，到脑机接口（BCIs），以及日益广泛应用于游戏、冥想和认知监测的消费级设备等一系列应用（Liu 等，2025；Portillo-Lara 等，2021）。低成本消费级脑电头戴设备的普及使这项技术走出实验室，进入日常环境，并推动了大规模脑电图数据采集（Jwa 和 Poldrack，2022；Yang 和 Jiang，2025）。

另一方面，这一转变引发了重大的隐私担忧，因为先前研究已表明，脑电信号可用于实施各种推断攻击，例如推断身份、性别、年龄、情绪和认知状态，以及是否患有某些疾病（Meng et al., 2024; Mandal and Saxena, 2022）。成员推断攻击的一种类型是成员推断

攻击（成员推理攻击），它使攻击者能够判断某个特定受试者的数据是否被用于训练深度学习模型（Shokri 等人, 2017）。成员推断攻击对脑电图模型构成重大威胁；例如，推断某人的脑电图曾被用于训练癫痫或抑郁症筛查模型，就可能泄露此人患有癫痫或抑郁症。这使得了解脑电图模型在何时以及为何容易受到成员推断攻击，以及哪些因素会导致脑电图模型存在成员推断攻击脆弱性，变得十分重要。

本文对影响脑电图模型易受成员推断攻击影响的因素进行了系统性实证研究。据我们所知，该领域的开创性研究是 (Cobileanet 等人, 2025) 的工作，该研究初步证明，基于卷积神经网络的脑电模型容易受到损失攻击。我们以这项工作为起点，通过考察三个代表不同采集设置和分类任务的脑电图数据集、六种卷积神经网络架构、一种可控的卷积神经网络架构（TestNetwork），以及

四种成员推断攻击方法。

除了衡量攻击成功率外，我们的核心目标是解释脑电图模型中成员信息泄露何时以及为何发生。对此，我们的主要发现和贡献如下：

- 我们使用 MoBI、驾驶员困倦和阿尔茨海默病数据集，对针对脑电图模型的成员推断攻击进行了全面评估。我们使用曲线下面积以及假正率为 1% 时的真正率指标，评估了损失、置信度、修正熵和离线似然比攻击。损失、置信度和修正熵攻击的表现相近，而 LiRA 会造成更严重的信息泄露。
- 我们发现，模型泛化与成员推断攻击脆弱性密切相关。训练集与测试集之间较大的损失差距和准确率差距始终对应着较高的攻击AUC。此外，延长模型训练时长会同时扩大泛化差距并加剧成员推断攻击脆弱性，这表明过拟合是导致成员推断攻击脆弱性的一个关键因素。
- 通过区分参与者数量与每位参与者贡献的脑电图数据量，我们的实验表明，增加不同参与者的数量通常会降低攻击成功率，而改变每位参与者的脑电图数据量所产生的影响则相对较弱。
- 我们考察模型规模和脑电图信号特征的影响。成员推断攻击脆弱性通常会随着模型规模增大而上升，但在模型规模较大时开始趋于饱和。相比之下，在不同脑电图数据集中，没有任何单一的头皮区域、频带、通道间相关模式，或时域与频谱表示能够一致地解释成员信息泄露。
- 我们通过受试者识别和留一受试者实验，探究个体特异性的作用。我们发现，在未见主体上评估时任务性能下降幅度更大的模型，也表现出更高的成员推断攻击脆弱性。这表明，对受试者特定表示的依赖是脑电图模型中成员信息泄露的重要来源。

总体而言，我们的研究结果表明，脑电图模型中的成员推断攻击脆弱性主要与泛化能力不佳、过拟合以及对受试者特定表示的依赖有关，而不是与脑电图信号的某个单一空间或频谱成分有关。这些结果为评估隐私风险以及设计更不易受到成员推断攻击影响的脑电图系统提供了实用指导。

2 相关工作

Shokri 等人（2017）首次提出了成员推断攻击，他们表明，拥有分类模型黑盒访问权限的攻击者可以判断给定样本是否属于该模型的训练集。Yeom 等人（2018）的后续研究确立了过拟合与成员推断攻击脆弱性之间的联系，并提出了用于成员推断的**损失**攻击。Song 和 Mittal（2021）指出，攻击准确率等聚合概念可能会低估成员推断攻击风险，并引入了基于指标的攻击。Carlini等人，2022提出，成员推断攻击应通过低假正率下的真正率（TPR）进行评估，而不是通过平均准确率或曲线下面积（AUC）进行评估。他们提出了似然比攻击（LiRA），该方法估计样本被纳入训练和未被纳入训练时模型输出的分布，并通过逐样本似然比对成员身份进行评分。Hu 等人（2022）和 Bai 等人（2024）提供了涵盖集中式环境和联邦式环境中成员推断攻击的综述。

尽管成员推断攻击已在图像、表格和文本数据领域得到广泛研究，但直到最近才开始应用于脑电图领域。据我们所知，(Cobilean 等，2025)首次研究了针对脑电图模型的成员推断攻击，并报告称脑电图模型容易受到损失攻击。我们以(Cobilean 等，2025)为起点，构建了一个更广泛的实验设置，涵盖多个数据集和成员推断攻击方法。Khanam et al., 2026)的近期研究也指出了联邦脑电图学习中的成员推理风险。

除成员推断攻击外，已有多项研究报告了脑电图系统中的隐私问题。研究表明，脑电记录会被动编码用户并不打算分享的多种属性，包括性别、年龄以及认知或情绪状态（Meng et al., 2024；Mandal and Saxena, 2022）。虽然这些攻击值得关注，但它们不同于本文重点讨论的成员推断攻击。多项研究也报告了脑电图的个体特异性（即脑电信号因人而异），例如（Campisi and La Rocca, 2014；Demuru and Fraschini, 2020）。这促使我们在第4.6节分析成员推理攻击脆弱性与个体特异性之间的联系。

3 方法

3.1 问题设定

令$\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ 表示由脑电图片段组成的训练数据集（我们交替使用片段和

样本这两个术语），其中$x_i \in \mathbb{R}^{C\times T}$表示包含$C$个通道和$T$个时间戳的脑电图片段，而$y_i \in \{1,\dots,K\}$表示其对应的类别标签。分类模型$f_\theta$由θ参数化，并在$\mathcal{D}_{\text{train}}$上训练，以预测未见测试样本的类别标签。我们用$\ell_\theta(x,y)$表示模型在样本$(x,y)$上的损失。对于分类任务，给定输入$x$，模型会生成一个对数几率向量：

$$\mathbf{z} = f_\theta(x) \in \mathbb{R}^K \tag{1}$$

可以使用softmax函数将logits转换为类别概率：

$$p_{\theta,k}(x) = \text{softmax}(\mathbf{z})_k = \frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)} \tag{2}$$

因此，预测类别为：

$$\hat{y}_\theta(x) = \underset{k\in\{1,\dots,K\}}{\arg\max}\ p_{\theta,k}(x) \tag{3}$$

3.2 数据集与预处理

我们在本研究中使用三个数据集：MoBI、驾驶员困倦和阿尔茨海默病。下文将简要说明这些数据集的获取方式，以及在模型训练前所进行的预处理。

多模态移动脑身成像（MoBI）数据集包含在儿童睁眼静息和创意电子游戏过程中采集的脑电图和头部运动记录（Ravindran等，2017）。脑电图以220 Hz的采样率采集，使用四通道消费级干电极Muse头戴式设备（AF7、AF8、TP9、TP10；参考电极位于FPz附近），并使用每个通道10 Hz的头带状态信号进行电极接触质量控制。按照（Cobilean 等，2025）的做法，我们使用同时包含静息和电子游戏记录的11名参与者，并训练模型f_θ，以区分在静息和电子游戏过程中记录的脑电图片段。

我们的数据集预处理遵循原始数据集的方法（Ravindran 等，2019），并分别应用于静息和游戏片段。每个脑电图通道均使用四阶零相位巴特沃斯滤波器进行 1–35 Hz 带通滤波，从 220 Hz 下采样至 110 Hz，通过 MEEGkit 使用伪迹子空间重建（ASR）（Mullen 等，2015）进行清理，随后进行 ±4-SD z分数裁剪和归一化。模型输入采用 2.5 秒窗口，并系统地改变重叠率：0%、25%、50%、75% 和约 99.6%（步长为 1 个采样点）；最后一种设置与先前研究（Cobilean 等，2025）相同。当头带状态信号表明电极接触不良（状态 >3）的时长超过窗口时长的 30% 时，该窗口将被排除。

驾驶员嗜睡数据集包含在模拟驾驶过程中采集的驾驶员脑电图记录设置中，参与者需要保持直线驾驶轨迹，同时随机引入车道偏离事件（Cao等，2019）。随后，后续研究为这些记录标注了警觉和困倦状态（Cui等，2022）。原始记录包含以 500 Hz 采集的 30 通道脑电图。我们对脑电图记录进行预处理，使用 1 至 50 Hz 的滤波器滤波，并将采样率降至 128 Hz。按照（Cui等，2022）的做法，我们提取偏离事件发生前 3 秒的脑电图片段，并根据先前研究中描述的反应时间标准，将其标记为警觉或困倦。随后对数据进行类别平衡，得到 2022 个片段（1011 个警觉和 + 1011 个困倦）。每个样本包含 30 个脑电图通道，每个通道有 384 个时间点。模型旨在识别每个片段对应的是警觉状态还是困倦状态。

阿尔茨海默病数据集 包含 88 名参与者闭眼静息状态下的脑电图记录（Miltiadous 等，2023）。我们使用了 36 名阿尔茨海默病患者和 29 名健康对照者（认知正常者）的脑电图数据。脑电图记录来自头皮上的 19 个通道，这些通道按照国际10-20系统布置，采样率为 500 Hz。原始脑电图记录采集时使用了 0.5–45 Hz 带通滤波器。在我们的实验中，信号被重采样至 160 Hz，并划分为 1 秒窗口，每个通道包含 160 个时间点。这些固定长度片段被用作脑电图分类和成员推断实验的输入。分类模型旨在区分阿尔茨海默病患者与认知正常者的脑电片段。

3.3 模型架构

遵循 Cobilean 等（2025）的先前研究，我们使用六种卷积神经网络（CNN）架构作为目标模型。此外，在模型复杂度分析（第 4.3 节）中，我们引入了一种名为 TestNetwork 的可控 CNN 架构。所有模型在默认配置下的参数量级均为 10^4，与先前研究中使用的紧凑模型规模一致。我们使用二元交叉熵损失训练模型，采用 Adam 优化器，学习率为 1×10^{-3}，批次大小为 128，并按 70%-15%-15% 划分训练集、验证集和测试集。

ShallowConvNet 是一种浅层神经网络，包含两个卷积层，后接 ReLU 激活函数、平均池化层和全连接输出层。其第一个卷积层用于提取时间特征，第二个卷积层为空间层，用于学习脑电图通道之间的依赖关系（Schirrmeister 等，2017）。

EEGNet 是 ShallowConvNet 的深层对应模型。它的第一层执行标准二维卷积，后接一个深度卷积层，再通过逐点卷积合并各通道的输出（Lawhern 等人，2018年）。

EEGNet2TL 通过在时序特征提取器中增加一层，对 EEGNet 进行了改进，以使用不同分辨率提取时间特征（Cobilean 等人，2025）。

EEGInception 将 Inception 模块应用于脑电图分类。它由两个 Inception 块组成，每个块都包含多个分支，用不同长度的卷积核提取时间特征。所有分支的输出会拼接在一起，经过平均池化后再进行卷积，随后送入最终的分类层（Santamaria-Vazquez 等人，2020）。

TIDNet 由两个阶段组成：时序特征提取和空间特征提取。提取的特征被送入分类层，该层为全连接层（Kostas 和 Rudzicz，2020）。

Conv1DModel 是一种轻量级的一维卷积神经网络，由三个一维卷积层组成。第一层使用大小为 64 的卷积核，之后每一层的卷积核大小均为前一层的一半（Kiranyaz 等人，2021）。

TestNetwork 由一系列时间卷积层、一个空间卷积层和一个线性分类器组成。我们可以分别调整每层的时间滤波器数和层数，从而在保持架构其余部分不变的情况下，大幅改变模型规模（∼1.9K 至 ∼830K 个参数）。默认的 TestNetwork 是规模最小的版本，包含 1 个时间层，每层有 20 个滤波器。

3.4 成员推断攻击与攻击方法

给定一个脑电图样本（x，y）以及对训练完成的分类模型 θ 的黑盒访问，成员推断攻击（MIA）的目标是确定（x，y）$\in \mathcal{D}_{\text{train}}$，也就是说，该样本是否曾用于训练目标模型。实施成员推断攻击有多种方法。Cobilean 等，2025 的原始研究采用了**损失**攻击。本文使用以下四种攻击。

损失攻击基于这样一种直觉：θ 通常会在成员样本上的损失低于非成员样本，因为其训练目标是最小化 $\mathcal{D}_{\text{train}}$ 上的损失（Yeom 等人，2018）。该攻击计算 $\ell_\theta(x, y)$，并将其与决策阈值进行比较。损失较低的样本会被预测为成员。

置信度攻击基于这样一种直觉：t θ 对成员样本的分类置信度高于非成员，因为模型在训练期间见过这些成员。攻击会计算预测类别概率的最大值：

$$c_\theta(x) = \max_{k\in\{1,\dots,K\}} p_{\theta,k}(x) = p_{\theta,\hat{y}_\theta(x)}(x) \tag{4}$$

并将其与决策阈值进行比较。如果$c_\theta(x) \geq$ 该阈值，则将 (x, y) 预测为成员；否则预测为非成员。

修改熵攻击在考虑真实标签y 的同时，衡量模型预测的不确定性（Song 和 Mittal，2021）。对于给定样本 (x, y)，其修正预测熵的计算方式为：

$$\begin{aligned}\text{Mentr}_\theta(x,y) = &-\big(1-p_{\theta,y}(x)\big)\log p_{\theta,y}(x) \\ &-\sum_{\substack{k=1\\k\neq y}}^{K} p_{\theta,k}(x)\log\big(1-p_{\theta,k}(x)\big)\end{aligned} \tag{5}$$

正确类别的概率越高，该分数越低；错误类别的概率越高，该分数越高。由于预期成员样本会产生较低的修正熵，当 $\text{Mentr}_\theta(x, y) \leq \tau_y$ 时，该攻击会将 (x, y) 预测为成员，其中 τ_y 是类别特定决策阈值；否则预测为非成员。

似然比（LiRA）攻击用于估计未在候选样本（x，y）上训练的模型在该样本上的表现（Carlini等人，2022）（这就是 LiRA 的“离线”变体）。攻击者在不包含（x，y）的辅助数据集上训练多个影子模型（我们训练了 64 个影子模型），并计算经 logit 缩放的置信度：

$$\phi_\theta(x,y) = \log\left(\frac{p_{\theta,y}(x)}{1-p_{\theta,y}(x)}\right) \tag{6}$$

这些结果用于估计样本特定的外部分布$\mathcal{N}(\mu_{\text{out}}$（$x$，$y$） σ^2_{out}（x，y）），其中$\mathcal{N}$（·）表示正态分布。令 Φ（·）表示标准正态分布的累积分布函数，则 LiRA 分数为：

$$\text{LiRA}_\theta(x,y) = \Phi\left(\frac{\phi_\theta(x,y)-\mu_{\text{out}}(x,y)}{\sigma_{\text{out}}(x,y)}\right) \tag{7}$$

较高的分数表示，与未观察到（x，y）的模型相比，目标模型对正确类别赋予了异常高的置信度。因此，当 LiRA_θ（x，y） $\geq \tau_{\text{LiRA}}$ 时，攻击将（x，y）预测为成员，否则预测为非成员。

3.5 评估指标

我们使用由成员和非成员构成的平衡数据集来评估成员推断攻击。当

成员样本被预测为成员时，即为真阳性；非成员被预测为成员时，即为假阳性；非成员被预测为非成员时，即为真阴性；成员被预测为非成员时，即为假阴性。随后，我们使用两个指标来衡量成员推理攻击的有效性。

攻击曲线下面积：我们改变成员推断攻击的成员判定阈值，以获得受试者工作特征（ROC）曲线。曲线下面积对应于ROC曲线下的面积。曲线下面积 = 0.5等同于随机猜测（无泄露），而曲线下面积 = 1.0则表示完美攻击。我们使用曲线下面积作为主要指标，因为它广受采用，并且属于平均情况指标，即它更重视总体上区分成员和非成员，而不是关注边缘情况。

1%假正率下的真正率：根据Carlini等人，2022，我们报告在固定假正率（FPR）为1%时的真正率（被正确识别的成员所占的比例）。这是一种边缘情况指标：它量化了攻击者能够找到多少成员，同时几乎不会错误地将非成员预测为成员。

4 分析与讨论

在本节中，我们将介绍关于影响脑电图模型中成员推断攻击脆弱性的关键因素的实证分析。为了量化某一因素与成员推断攻击脆弱性之间的关联程度，我们通常会报告该因素与攻击成功率（主要使用曲线下面积指标进行衡量）之间的Pearson相关系数 r，以及用于检验统计显著性的p值，同时提供相关的可视化图表。系数$r \in [-1,1]$ 衡量某一因素与成员推断攻击脆弱性之间线性相关的强度和方向。接近 +1 或-1的值表示存在显著相关性，而接近0的值表示不存在关联。对于p值，我们将$p < 0.05$ 视为具有统计显著性。

4.1 模型泛化

我们发现，影响成员推断攻击脆弱性的一个关键因素，是模型对训练数据集的泛化能力（相对于过拟合程度）。为量化这一点，我们使用两个概念：训练测试**损失差距** 和**准确率差距**。令$\mathcal{D}_{\text{test}} = \{(x_j, y_j)\}_{j=1}^{M}$表示与$\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$不相交的测试数据集。平均训练损失和测试损失为：

$$L_{\text{train}} = \frac{1}{N}\sum_{i=1}^{N} \ell_\theta(x_i, y_i) \quad L_{\text{test}} = \frac{1}{M}\sum_{j=1}^{M} \ell_\theta(x_j, y_j) \tag{8}$$

训练测试损失差距定义为：

$$G_{\text{loss}} = L_{\text{test}} - L_{\text{train}} \tag{9}$$

较大的正G_{loss} 表示模型在训练样本上的损失远低于在未见样本上的损失，表明过拟合更严重。同样，训练准确率和测试准确率定义如下：

$$\begin{aligned} A_{\text{train}} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\hat{y}_\theta(x_i) = y_i] \\ A_{\text{test}} &= \frac{1}{M}\sum_{j=1}^{M} \mathbb{I}[\hat{y}_\theta(x_j) = y_j] \end{aligned} \tag{10}$$

其中 $\mathbb{I}[\cdot]$为指示函数。训练与测试准确率差定义如下：

$$G_{\text{acc}} = A_{\text{train}} - A_{\text{test}} \tag{11}$$

因此，G_{loss} 和G_{acc} 的正值越大，表示泛化能力越差。

攻击AUC与G_{loss} 和G_{acc} 的关系结果见图1。结果显示，攻击AUC与G_{loss} 和G_{acc} 之间存在正相关，且相关性高度显著（Pearson$r$ 介于0.59和0.80之间，p值低于0.001）。换言之，当模型在训练数据集与测试数据集上的表现不同时（例如，模型泛化能力较差），攻击AUC会升高。根据MoBI数据集，与其他（更常见的）重叠方式相比，采用步长为1的脑电图窗口划分时，这种关系略弱；不过，其余没有窗口重叠的数据集仍显示出显著的正相关（Driver-Drowsiness的Pearson$r$ 介于0.56和0.68之间；阿尔茨海默病数据集介于0.74和0.85之间）。

4.2 数据集特征

接下来，我们研究脑电图数据集的不同特征是否会影响成员推断攻击脆弱性。具体而言，我们限制纳入$\mathcal{D}_{\text{train}}$的数据中的参与者数量，然后训练模型，例如仅使用 1 名参与者、2 名参与者、3 名参与者，依此类推。攻击AUC如图2所示。从图中可以看出，随着参与者数量增加，攻击AUC总体呈下降趋势。这一趋势似乎适用于多个模型（Conv1DModel、EEGInception、EEGNet）。如图3所示，采用非重叠窗口划分时，阿尔茨海默病数据集的曲线下面积与MoBI数据集的趋势一致。

虽然上述发现与 Cobilean 等，2025 的先前研究一致，但 TPR @ 1% FPR指标的表现有所不同。如图

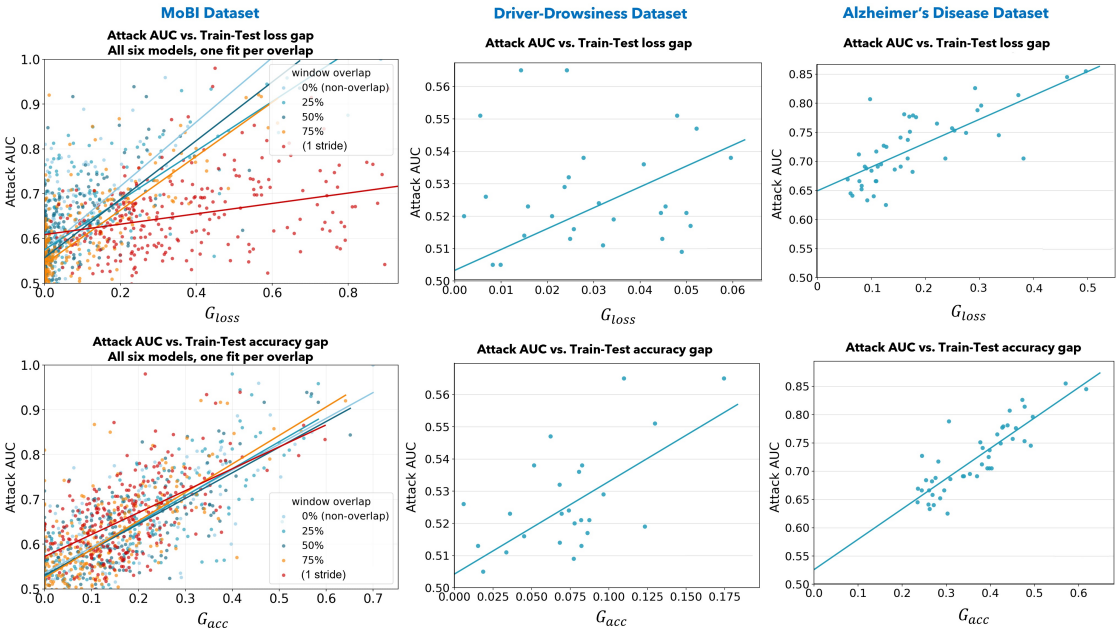

图1：**模型泛化结果**。第一行：攻击AUC与损失差距的关系；第二行：攻击AUC与准确率差距的关系。左侧一组对应 MoBI 数据集（EEG窗口划分有五种重叠比例），中间一组对应驾驶员嗜睡数据集，右侧一组对应阿尔茨海默病数据集（后两个数据集均采用无重叠窗口划分）。

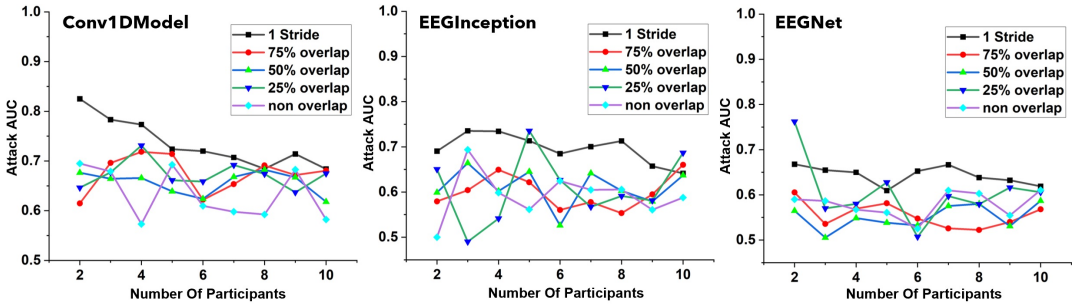

图2：**攻击AUC与参与者数量的关系**。尽管存在一些波动，增加参与者数量仍会降低攻击者的AUC（MoBI数据集）。

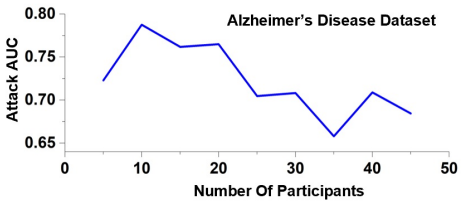

图3：**攻击AUC与参与者数量的关系**（阿尔茨海默病数据集，采用无重叠窗口划分）。结果与 MoBI 数据集相似，虽有波动，但总体呈下降趋势。

4 所示，随着参与者人数增加，真正率逐渐下降（从 0.18 降至最低 0.006）。在全部六个模型中，TPR @ 1% FPR 与参与者人数之间的相关性均为负，且具有统计学显著性（根据 p值判断）。

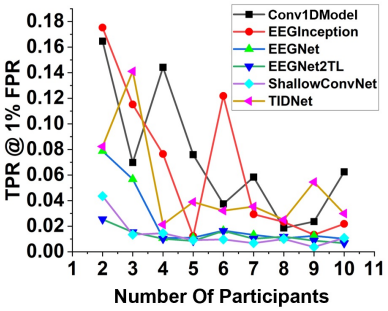

图4：**参与者数量对应的真正率（TPR）@1% 假正率（FPR）**。此实验仅针对 MoBI 数据集的 1 步长窗口重叠设置进行。

我们还进行了一项实验，固定参与者人数，只改变

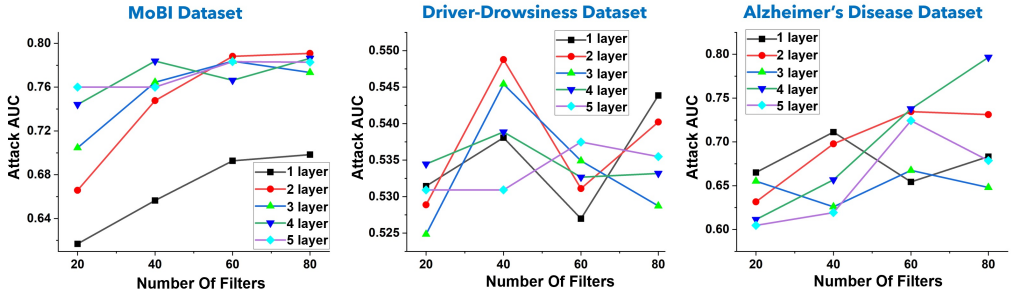

图5：**模型规模的影响（滤波器和层数）**。MoBI数据集（左）、驾驶员嗜睡数据集（中）和阿尔茨海默病数据集（右），采用无重叠窗口划分。

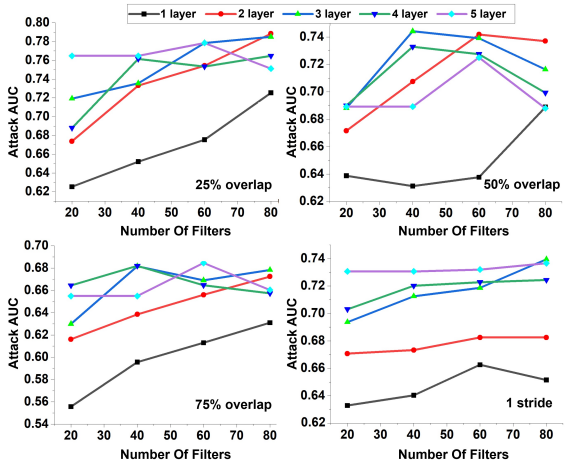

图6：在 MoBI 数据集上，对于可控的 TestNetwork，攻击 AUC 随每层滤波器数变化的情况，窗口重叠率分别为 25%、50%、75% 和 1。

训练窗口数量，即每位参与者有多少数据被纳入 $\mathcal{D}_{\text{train}}$。受篇幅限制，我们没有提供该实验的图表，但会以文字讨论结果。我们发现，训练数据集规模对成员推断攻击脆弱性的影响较弱。在 30 种模型和脑电窗口重叠条件下，训练数据集规模与攻击AUC之间的平均相关性为 -0.16，与 1%假正率下的真正率之间的平均相关性为 -0.11。根据 p值，这些相关性既不强，也不具有统计显著性。我们特别指出这一发现，因为在先前研究（Co-bilean 等人，2025）中，训练数据集规模与参与者人数相互混杂。在我们的实验中将两者分开后，我们发现训练数据集规模的影响相对较小，而参与者人数的影响则更为显著。

总体而言，这些结果表明，成员推理攻击的脆弱性由不同个体的数量所驱动。训练中的参与者越多，每位参与者就越不容易被记住，即越不易受到成员推理攻击。相比之下，每位参与者的数据量较少或较多，对成员推断攻击的影响并不显著。

的参与者并不会对成员推断攻击产生同样显著的影响。

4.3 模型与训练特征

在本节中，我们研究脑电图模型及其训练的特征（例如卷积神经网络的规模和复杂度）是否会影响成员推断攻击脆弱性，因为人们通常认为，更大的模型更容易过拟合和记忆化。首先，我们使用可控的 TestNetwork，改变层数 $\in \{1,2,3,4,5\}$ 和每层滤波器数$\in \{20,40,60,80\}$。该实验的结果如图5所示。我们观察到，总体而言，攻击AUC往往会随着层数和滤波器数的增加而上升。所得卷积神经网络中的参数数量与攻击AUC之间呈显著正相关（$r = +0.34$ $p < 0.001$）。然而，当参数数量超过约 10^5 后，这种影响便趋于饱和。这些发现与以下直觉一致：参数数量更多的大型模型更容易受到记忆化和成员推断攻击的影响。我们还注意到，这种影响在 MoBI 和阿尔茨海默病数据集上更明显，而在 Driver-Drowsiness 数据集上相对较弱，这表明数据集也可能发挥作用。

此外，我们进行图6所示的实验，以分析MoBI数据集预处理中的窗口重叠量是否会影响上述发现。我们报告了窗口重叠分别为25%、50%、75%和1步长重叠时的结果。总体而言，结果支持此前的发现，即增加滤波器数量会提高攻击AUC。然而，1步长重叠实验显示出最弱的相关性，因为连续窗口之间存在较大的重叠部分。此外，将层数从1层增加到2层似乎会产生显著影响，而其余层数（2、3、4、5）通常可以获得相近的曲线下面积。

接下来，我们分析训练轮数（即训练时长）对攻击AUC和训练测试损失差距G_{loss}的影响。结果见图

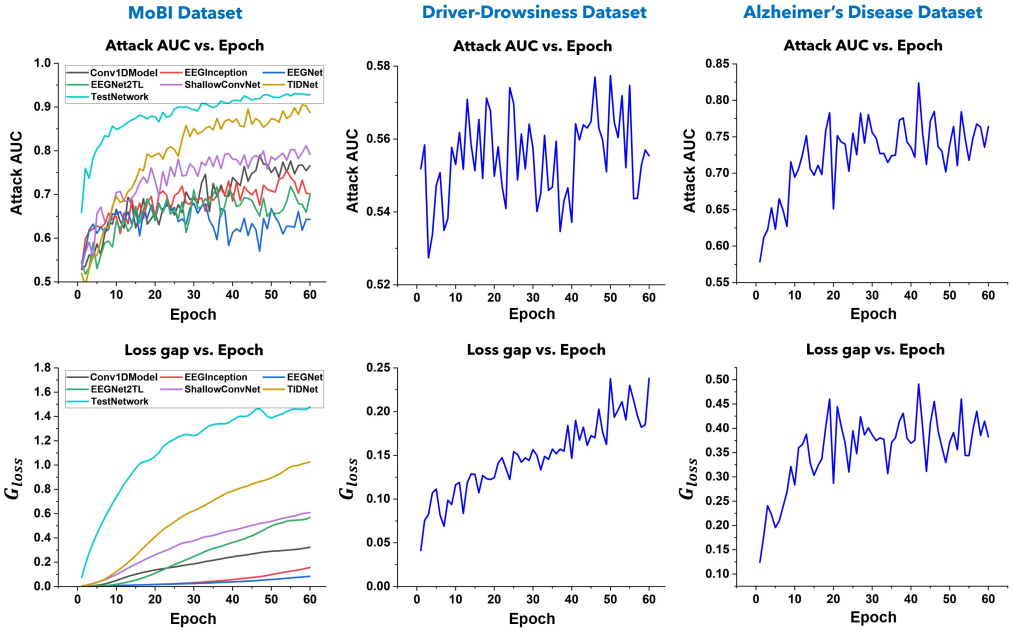

图7：**模型训练所用训练轮数与攻击AUC及训练测试损失差距的关系**。从左到右依次为 MoBI、驾驶员困倦和阿尔茨海默病数据集。结果表明，随着训练轮数增加，攻击AUC和训练测试损失差距也随之增大。

7。我们观察到，攻击AUC会随着训练轮数增加而上升（各模型的r 从 +0.24升至 +0.75，依据p值判断均具有统计显著性）。我们还观察到G_{loss} 也在增大，这意味着训练数据集与测试数据集之间的损失差异在扩大。训练轮数与AUC和G_{loss}均呈正相关，这很直观：随着模型在$\mathcal{D}_{\text{train}}$上不断训练，它开始对$\mathcal{D}_{\text{train}}$过拟合，从而扩大损失差距并增加成员推断攻击脆弱性。

4.4 脑电信号的空间和频谱预测因子

接下来，我们考察成员推断攻击脆弱性是否可归因于脑电信号的某种特定空间、频谱或表征属性。在这些分析中，没有任何单一的数据特征始终表现为攻击成功率的主导预测因子。下面详细介绍我们的分析。

脑电通道区域：图 8 比较了三个数据集中不同脑电通道子集的攻击曲线下面积。在 MoBI 中，额部（AF7-AF8）、颞部（TP9-TP10）、半球以及单通道配置的攻击性能总体相近，曲线下面积通常约为 0.55 至 0.65。同样，驾驶员嗜睡数据集中的区域差异很小，所有通道子集的曲线下面积均处于约 0.50-0.60 的狭窄范围内。阿尔茨海默病数据集中的区域差异较大，其中完整电极阵列的攻击曲线下面积最高，而枕部通道的曲线下面积低于其他通道组。不过，这种区域模式在不同数据集之间并不一致。总体而言，这些结果表明，成员身份易受攻击的脆弱性并不局限于特定头皮区域；与某些受限子集相比，使用所有可用通道可获得更高的攻击曲线下面积。

频带：接下来，我们将输入限制为单个频带（图9）。在 MoBI 数据集上，θ波、α波、β波和γ波的攻击表现几乎相同（Kruskal-Wallis检验$p = 0.84$；曲线下面积差值为 = 0.005），每个单独频带的曲线下面积约为 0.59，而宽带信号为 0.612（全部$p < 0.001$）。值得注意的是，所有单独频带的攻击表现仍显著高于随机猜测。其他数据集并未完全呈现相同模式：驾驶员嗜睡数据集中各频带之间差异很小，曲线下面积仍接近 0.5；而在阿尔茨海默病数据集中，若干单独频带的曲线下面积与宽带条件相当或更高。总体而言，这些结果表明，在不同数据集中，没有任何特定频带能够始终预测成员推断攻击脆弱性。

输入表示：将时域

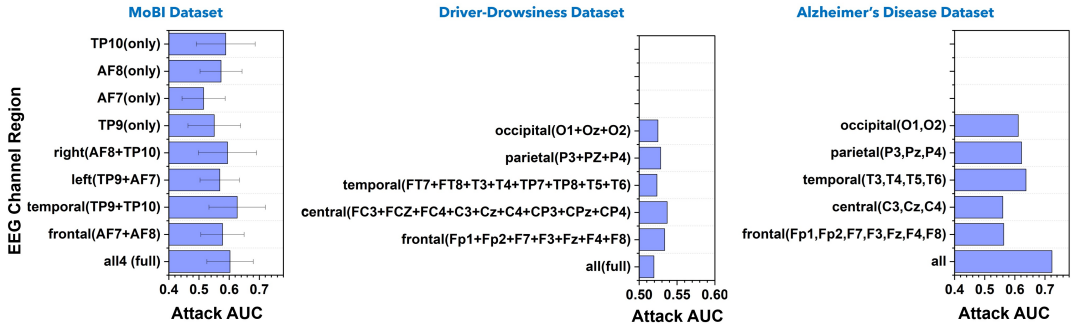

图8：**EEG 通道区域**。三个数据集按通道子集划分的攻击AUC。没有任何单一区域始终占优势；整套装置的表现与任何子集相当或更好。

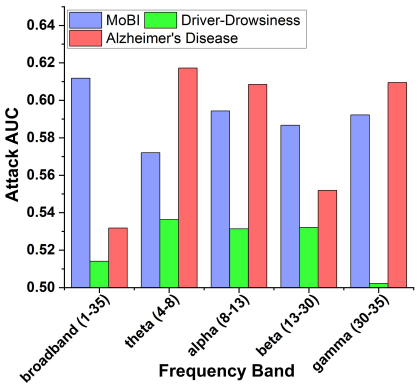

图9：**频带**。脑电图输入分别限制在 θ波、α波、β波或γ波频带时的攻击AUC，并与完整宽带信号进行比较。

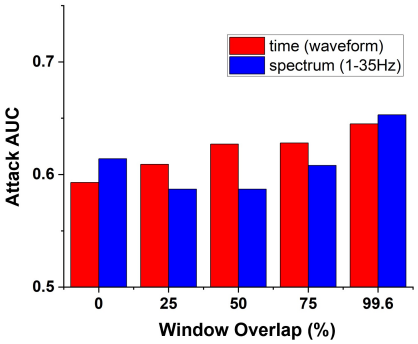

图10：**频谱表示**。MoBI数据集上时域波形（红色）与幅度频谱（1–35 Hz，蓝色）输入表示的攻击AUC。两种表示在每种重叠情况下均产生相当的攻击强度，没有一种始终占优。

脑电图波形替换为其 1–35 Hz 幅度谱，对易受攻击程度的影响也很小（图10）。在五种 MoBI 窗口重叠条件下，两种表示的攻击性能相当：平均曲线下面积为 0.621，

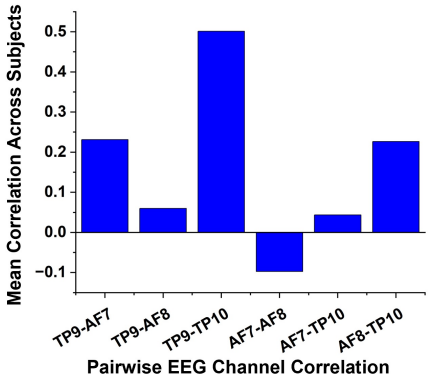

图11：**通道间相关性**。MoBI数据集的平均相关性与通道对的关系。

波形为 0.610，频谱表示为 0.610，两者之间没有显著差异（$p = 0.40$）。在不重叠和步长为 1 的情况下，频谱的表现略好；而在中间重叠程度下，波形的表现略好。将频谱表示限制在单个频带后，攻击表现相较于完整频谱再次下降（$p = 0.0002$）。这些发现表明，可以从时域和频谱表示中获取成员信息，而非仅取决于特定输入域。

通道间相关性： 最后，我们考察通道间相关性是否与成员推断攻击脆弱性相关。结果如图11所示。在单样本步长（1-stride）下，平均通道间相关性与攻击性能仅呈现微弱且不显著的关系（$r = 0.16$, $p = 0.65$）。在较高的窗口重叠率下，使用六个成对相关性值的成员推断攻击取得了约0.54的曲线下面积，这表明相关性结构包含一些与成员身份相关的信息，但其预测能力较弱。如图11所示，最强的相关性出现在两个颞区电极TP9-TP10之间。

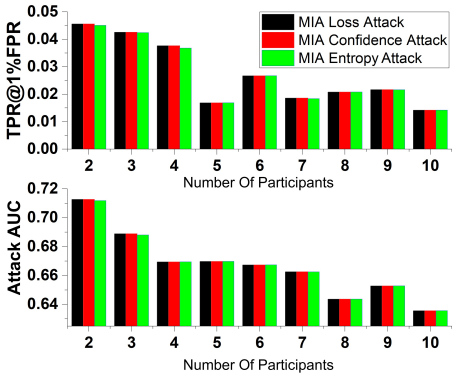

图12：MoBI数据集上损失攻击、置信度攻击和修改熵攻击的曲线下面积和1%假正率下的真正率结果。

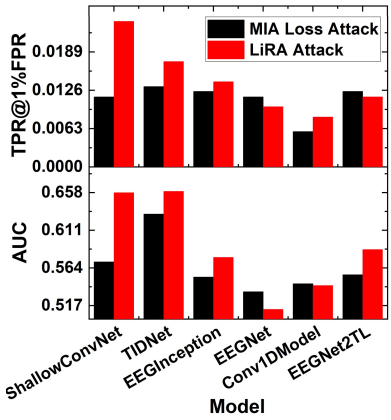

图13：LiRA 与损失攻击在 MoBI数据集上的攻击AUC 和 1%假正率下的真正率对比。

4.5 成员推理攻击方法的选择

如第3.4节所述，可以通过多种方法执行成员推理攻击，例如损失、置信度、修改后的熵以及LiRA 攻击方法。我们的结果默认采用损失攻击，与 Cobilean 等，2025一致。本节将探讨不同攻击的影响。

为确认我们的发现和结论不依赖于特定的攻击方法，我们首先比较了损失攻击、置信度攻击和修改熵攻击。我们在全部六种模型架构、三个数据集以及MoBI数据集的多种窗口重叠设置下进行了比较。总体而言，我们观察到这三种攻击的曲线下面积非常相近（最大差异为0.02）。由于篇幅限制，本文不报告全部结果，而仅将图12作为代表性结果。鉴于三种攻击方法的性能高度相似，我们得出结论：损失攻击具有良好的代表性，适合在全文中作为主要攻击方法。我们注意到，这些攻击之间的高度相似性是由于所有数据集都执行二分类，即类别数量为K= 2。在多分类情况下，攻击的表现可能会大不相同。

另一个值得考虑的问题是，我们的 LiRA 攻击与损失攻击相比表现如何。我们在图13中报告了这项实验的结果。我们发现，LiRA 的表现通常优于损失攻击：平均曲线下面积从 0.570 上升至 0.606，1%假正率下的真正率则提高了约 3 倍。对于成员推断攻击脆弱性较高的架构，这种提升最为显著；而对于脆弱性最低的架构（例如 EEGNet），提升则微乎其微。这表明，选择不同的成员推理攻击方法会放大现有的成员推断攻击脆弱性。不过，这一结果在实践中的一个意义是，如果只关注

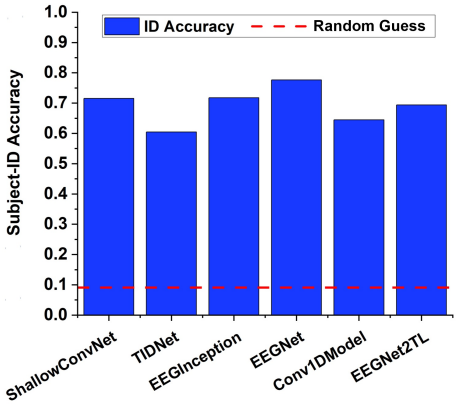

图14：MoBI数据集上的**受试者识别准确率** 实验结果。

损失攻击，可能会低估最坏情况下的成员推理攻击风险。未来可以尝试提出效果更好的成员推断攻击，这将是一个有价值的研究方向，因为更换攻击方法似乎还能进一步提升攻击效果。

4.6 能否识别受试者?

为了检验脑电信号的成员推理攻击脆弱性与个体特异性之间的关系，我们测试是否可以利用参与者的脑电信号识别数据集中的受试者（参与者）。我们通过将最后的决策层替换为 N 类分类头，训练卷积神经网络模型架构，以预测给定脑电窗口属于哪个受试者（即哪个参与者）。模型的受试者识别准确率结果见图14。鉴于我们的 MoBI数据集包含11名参与者，随机猜测的受试者识别准确率接近0.091。从图14可以观察到，所有架构的表现都显著优于随机猜测，通常达到随机猜测准确率的6—7倍。

尤其是 EEGNet 的表现非常出色。基于这些结果，我们得出结论：脑电窗口确实包含个体特异性特征，并且可以利用参与者的脑电图识别参与者。

最后，我们回到原始的静息态与游戏分类任务，比较模型在两种设置下的准确率：（i）在 $\mathcal{D}_{\text{train}}$ 和 $\mathcal{D}_{\text{test}}$ 中使用相同主体的数据；（ii）采用留一主体（LOSO）训练，此时同一主体的数据不能同时出现在 $\mathcal{D}_{\text{train}}$ 和 $\mathcal{D}_{\text{test}}$ 中。结果如图15所示。我们观察到，与第一种设置相比，LOSO 设置下静息态与游戏任务的准确率有所下降。这表明，如果模型在训练期间见过相同主体的数据，那么它在测试数据上的表现会更好，这进一步印证了我们此前关于个体特异性的发现。此外，图15中的第二幅图显示，MIA 曲线下面积与主体内准确率和留一主体准确率之差呈正相关（但这种相关性在统计上并不显著）。这意味着，对未见主体的泛化能力较弱的模型（即主体内准确率与留一主体准确率之差较大）更容易受到成员推断攻击（即 MIA 曲线下面积较高）。结合此前的结果，这些发现表明，对脑电图个体特异性的依赖是成员信息泄露的驱动因素之一：从训练期间见到相同主体的数据中获益最多的模型，也表现出更高的成员推断攻击脆弱性。

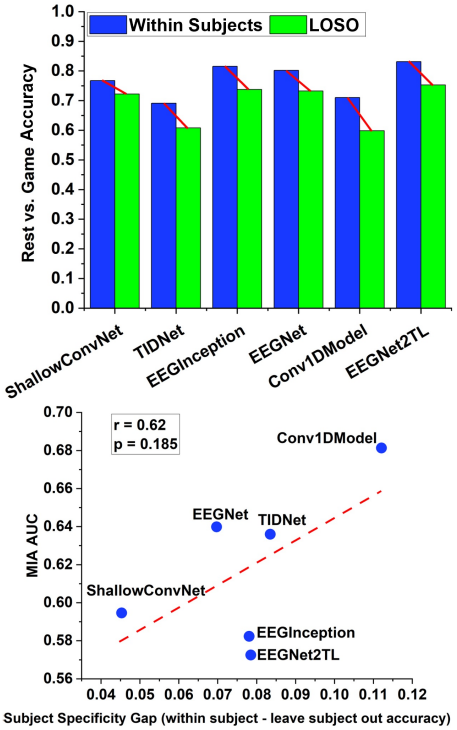

图15：MoBI数据集在受试者内训练和留一受试者法训练设置下，静息态与游戏分类任务的准确率。

5 结论

本文研究了多个数据集、卷积神经网络架构和成员推断攻击方法中，影响脑电图模型成员推断攻击脆弱性的因素。结果表明，模型泛化能力较差（即训练集与测试集之间的损失和准确率差距较大）与更高的成员推断攻击脆弱性密切相关。增加参与者数量通常会降低攻击成功率，而每位参与者贡献的脑电图数据量影响相对较弱。模型容量和训练时长也会增加脆弱性，而特定头皮区域、频段或输入表示均无法持续解释成员信息泄露。此外，我们的受试者识别和留一受试者法实验表明，依赖主体特定脑电表示会导致成员推断攻击脆弱性。

未来工作有多个方向。首先，我们的攻击针对脑电窗口（即片段）实施成员推理攻击；未来工作也可以考虑以受试者为单位的攻击。其次，我们使用不同的架构训练卷积神经网络模型，未来工作可以研究脑电基础模型。最后，在脑电图场景中防御成员推断攻击也是一个重要的未来研究方向。

致谢

本研究获得土耳其科学与技术研究理事会（T¨urkiye）（T¨UB˙I- TAK）的资助，项目编号为123C595；同时还获得KUIS人工智能主题种子研究基金和BAGEP杰出青年科学家奖的支持。作者感谢T¨UB˙ITAK、KUIS AI和科学学院提供的支持。

AI 使用披露：我们使用 Claude（Anthropic）作为 AI 助手，协助修改文本，并调试和记录实验中使用的 Python 代码。所有实验均由作者设计并执行，所有报告的结果、图表及其解读均为作者原创。作者对论文内容承担全部责任。

参考文献

Bai, L., Hu,H., Ye,Q., Li, H., Wang,L., and Xu, J.（2024年）。联邦-

学习中的成员推断攻击与防御：综述。*ACM*计算调查，57(4):1–35。Campisi, P. and La Rocca, D.（2014）。用于自动基于生物特征的用户识别的脑电波。*IEEE*信息取证与安全汇刊，9(5):782–800。Cao, Z., Chuang, C.-H., King, J.-K., and Lin, C.-T.（2019）。持续注意力驾驶任务期间的多通道脑电图记录。*ScientificData*，6(1):19。Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F.（2022）。从第一性原理出发的成员推断攻击。载于*2022 IEEE* 安全与隐私研讨会（*SP*），第1897–1914页。IEEE。Cobilean, V., Mavikumbure, H. S., Drake, D., Stuart, M., and Manic, M.（2025）。研究针对脑机接口系统卷积神经网络模型的成员推断攻击。*IEEE*生物医学与健康信息学杂志。Cui, J., Lan, Z., Liu, Y., Li, R., Li, F., Sourina, O., and M¨uller-Wittig, W.（2022）。一种用于从单通道脑电图进行跨被试驾驶员疲劳检测的紧凑且可解释的卷积神经网络。*Methods*，202:173–184。Demuru, M. and Fraschini, M.（2020）。脑电图指纹识别：基于功率谱非周期成分的被试特异性特征。*Computers in Biology and Medicine*，120:103748。Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang, X.（2022）。针对机器学习的成员推断攻击：综述。*ACM*计算调查，54(11s):1–37。Jwa, A. S. and Poldrack, R. A.（2022）。应对神经科学数据中的隐私风险：从数据保护到伤害预防。*Journal of Law and the Biosciences*，9(2)。Khanam, T., Siuly, S., Wang, K., Whittaker, F., and Wang, H.（2026）。评估联邦脑电图学习中的多级成员推断风险。*Brain Informatics*。Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J.（2021）。一维卷积神经网络及其应用：综述。*Mechanical SystemsandSignal Processing*，151:107398。Kostas, D. and Rudzicz, F.（2020）。思考者不变性：让深度神经网络能够跨越更多人应用于脑机接口。*Journal of Neural Engineering*，17(5):056008。Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J.（2018）。EEGNet：用于基于脑电图的脑机接口的紧凑型卷积神经网络。*Journal of Neural Engineering*，15(5):056013。Liu, X.-Y., Wang, W.-L., Liu, M., Chen, M.-Y., Pereira, T., Doda, D. Y., Ke, Y.-F., Wang, S.-Y., Wen, D., Tong, X.-G., et al.（2025）。基于脑电图的脑机接口在医学领域的最新应用。*Military Medical Research*，12(1):14。Mandal, A. and Saxena, N.（2022）。SOK：你的思想透露了许多关于你的信息：论通过脑电波设备造成的隐私泄露。载于*Proceedingsofthe 15thACM Conferenceon Security and Privacy in Wireless and Mobile Net- works*，第175–187页。Meng, L.、Jiang, X.、Jia, T. 和 Wu, D.（2024）。在基于脑电图的脑机接口中同时保护多种类型的隐私。载于 *2024 IEEE* 国际系统、人和控制论会议（*SMC*），第2894–2899 页。IEEE。Miltiadous, A.、Tzimourta, K. D.、Afrantou, T.、Ioannidis, P.、Grigoriadis, N.、Tsalikakis, D. G.、Angelidis, P.、Tsipouras, M. G.、Glavas, E.、Giannakeas, N. 等（2023）。来自常规脑电图的阿尔茨海默病、额颞叶痴呆和健康受试者头皮脑电记录数据集。*Data*，8(6):95。Mullen, T. R.、Kothe, C. A.、Chi, Y. M.、Ojeda, A.、Kerth, T.、Makeig, S.、Jung, T.-P. 和 Cauwenberghs, G.（2015）。使用可穿戴干式脑电图进行实时神经影像和认知监测。*IEEE* 生物医学工程汇刊，62(11):2553–2567。Portillo-Lara, R.、Tahirbegi, B.、Chapman, C. A.、Goding, J. A. 和 Green, R. A.（2021）。关注差距：脑电图脑机接口的最先进技术与应用。*APLBioengineering*，5(3)。Ravindran, A. S.、Cruz-Garza, J.、Kopteva, A.、Paek, A.、Mobiny, A.、Hernandez, Z. 和 Contreras-Vidal, J.（2017）。多模态移动脑身成像（MoBI）数据集：用于评估儿童在创意电子游戏过程中相关的神经和头部运动反应。*IEEE Dataport*。Ravindran, A. S.、Mobiny, A.、Cruz-Garza, J. G.、Paek, A.、Kopteva, A. 和 Contreras Vidal, J. L.（2019）。评估儿童在公共空间进行电子游戏时的神经活动：一种深度学习方法。*Journal of Neural Engineering*，16(3):036028。Santamaria-Vazquez, E.、Martinez-Cagigal, V.、Vaquerizo-Villar, F. 和 Hornero, R.（2020）。EEG-Inception：一种用于辅助性基于事件相关电位的脑机接口的新型深度卷积神经网络。*IEEE* 神经系统与康复工程汇刊，28(12):2773–2782。Schirrmeister, R. T.、Springenberg, J. T.、Fiederer, L. D. J.、Glasstetter, M.、Eggensperger, K.、Tangermann, M.、Hutter, F.、Burgard, W. 和 Ball, T.（2017）。使用卷积神经网络进行脑电解码和可视化的深度学习。*HumanBrain Mapping*，38(11):5391–5420。Shokri, R.、Stronati, M.、Song, C. 和 Shmatikov, V.（2017）。针对机器学习模型的成员推断攻击。载于 *2017 IEEE* 安全与隐私研讨会（*SP*），第 3–18 页。IEEE。Song, L. 和 Mittal, P.（2021）。系统评估机器学习模型的隐私风险。载于 第 *30* 届 *USENIX* 安全研讨会（*USENIX Security 21*），第 2615–2632 页。Yang, H. 和 Jiang, L.（2025）。脑机接口时代的神经数据处理监管：伦理问题与法律方法。*DigitalHealth*，11。Yeom, S.、Giacomelli, I.、Fredrikson, M. 和 Jha, S.（2018）。机器学习中的隐私风险：分析其与过拟合的联系。载于 *2018 IEEE* 第 *31* 届计算机安全基础研讨会（*CSF*），第 268–282 页。IEEE。