脑机接口对抗安全研究综述

孟璐斌[1] 王宇迪[2] 罗 飞[2] 伍冬睿[1]
1. 人工智能与自动化学院 华中科技大学 武汉 430074;
2. 东风汽车集团有限公司研发总院 武汉 430058

摘 要 脑机接口作为人机交互的重要技术路径，在医疗康复、无障碍通信、神经调控等多个领域展现出广阔的应用前景。然而，脑机接口系统高度依赖机器学习对脑电信号进行解码，近年来暴露的对抗攻击问题使得其安全性面临严峻挑战。对抗攻击通过在输入信号中添加微小扰动，即可诱导模型产生错误预测，进而可能导致设备失控或被远程操控，严重威胁系统可靠性与用户人身安全。本文聚焦非侵入式脑机接口中的对抗安全问题，系统梳理当前研究进展，从攻击与防御两个维度展开综述。通过全面评估现有成果，本文总结了当前脑机接口系统在安全性方面存在的关键挑战，并提出未来值得关注的研究方向，为构建安全、可靠、可部署的脑机接口系统提供理论支持与技术参考。

关键词 脑机接口；对抗攻击；对抗防御

A Survey on Adversarial Security in Brain-Computer Interfaces

Meng Lu-bin[1], Wang Yu-di[2], Luo Fei[2], Wu Dong-rui[1]
1.School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, Hubei province, China;
2. Dongfeng Motor Corporation Research Center, Wuhan 430058, Hubei province, China）

Abstract: Brain-computer interfaces (BCIs) have shown great potential in rehabilitation, assistive communication, and neuromodulation. However, their strong reliance on machine learning for EEG decoding exposes them to adversarial attacks. These attacks add subtle perturbations to input signals, which can mislead models and cause incorrect predictions, leading to system malfunction or even remote manipulation. This paper focuses on adversarial security in non-invasive BCIs, providing a systematic review of current research on both attacks and defences. It summarises key challenges and outlines future directions to support the development of secure and reliable BCI systems.

Keywords: Brain-Computer Interface; Adversarial Attack; Adversarial Defense

脑机接口（Brain−Computer Interface, BCI）旨在搭建大脑与外部设备（如假肢、轮椅、机器人、计算机等）之间的直接交互通路，通过解读神经活动，实现对人体感知与运动功能的替代、恢复、增强或补充[1]。该技术已在医疗康复、神经调控、人机交互等多个前沿领域展现出广阔的应用前景[2-5]。

非侵入式BCI系统[6]一般包含五个关键模块，如图1所示：信号采集、信号处理、特征提取、模式识别与外部控制。其中，信号采集通常通过电极阵列从头皮记录脑电（Electroencephalography，EEG）信号，经预处理后传入特征提取模块以捕获任务相关的关键信息。随后，模式识别模块通过分类或回归手段对这些特征进行解码，从而推断用户意图，并将结果转化为控制指令反馈至外部设备。在整个流程中，特征提取与模式识别是决定BCI系统性能的核心环节。近年来，随着机器学习的不断发展，其在这两个模块中的应用日益广泛，并显著提升了BCI的解码精度[7-13]。

尽管机器学习技术为BCI系统带来诸多性能突破，但其安全性问题却逐渐引起关注，特别是对抗攻击带来的潜在威胁[14-17]。该类攻击最早出现在图像识别领域，指攻击者向输入中添加人眼难以察觉的微弱扰动，即可误导模型产生错误预测。例如，通过在图像中嵌入细微噪声，原本被正确识别为“猫”的样本可能被错误分类为“狗”。此类攻击方式已被证实在语音识别、文本分析、生物识别等多个领域同样有效，引发了学术界

*[基金项目] 浙江大学脑机智能全国重点实验室开放课题（BMI2400015）.

对机器学习模型鲁棒性的担忧。

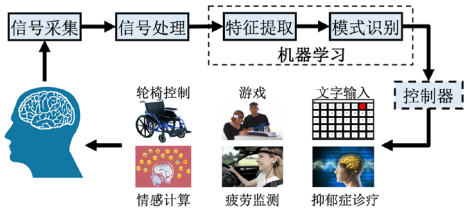


图1 非侵入式脑机接口闭环流程图

Fig.1 A non-invasive EEG-based BCI system

BCI系统同样存在对抗攻击的安全隐患。由于BCI系统高度依赖模型对脑电信号的精准解码，攻击者若在输入中注入精心设计的扰动，即便在统计特性或可视化层面几乎不可见，也可能使系统输出错误决策。例如，恶意扰动可能诱导脑控义肢执行偏离意图的动作，或令脑控轮椅误判用户指令，从而引发严重的人身安全风险。更为严重的是，若攻击者在模型训练阶段植入后门机制，未来可通过特定输入远程操控系统行为，造成不可控的安全后果。

此外，BCI系统的物理开放性进一步放大了其易受攻击的风险。由于EEG信号采集环境通常缺乏物理隔离，电极直接暴露在外，极易受到电磁干扰等物理信号的影响。同时，一些BCI设备采用无线传输或开放接口，也为硬件层级的攻击提供了可能，使得从源头干扰数据成为现实，且后续算法难以完全修正。

本文聚焦于非侵入式BCI系统中的对抗安全问题，围绕攻击与防御两大维度展开系统综述。在攻击层面，重点介绍逃逸攻击、污染攻击与物理域攻击三个研究方向，剖析其实现机制与代表性研究进展；在防御层面，归纳主流方法，包括基于检测、鲁棒性训练与输入变换的防御方法。最后，本文总结当前BCI系统在对抗安全方面所面临的关键挑战，并提出值得关注的未来研究方向，旨在为BCI系统的安全可控设计提供理论基础与工程指导。

1 对抗攻击

1.1 对抗攻击简介

对抗攻击是一种通过向输入样本中添加难以察觉的微小扰动，从而误导机器学习模型输出错误决策的技术手段。尽管这些扰动在感知层面几乎不可分辨，却可能导致模型产生严重的预测偏差。该现象最早在图像识别任务中被发现，例如对图像像素进行轻微修改，即可使模型将原本正确分类为“熊猫”的图像误判为“长臂猿”，而人眼几乎察觉不到图像的变化。近年来，对抗攻击的研究不断扩展至语音识别、自然语言处理、自动驾驶、医疗影像等多个领域，其潜在的安全隐患引发广泛关注。

根据攻击所处阶段的不同，对抗攻击通常分为以下两类。

逃逸攻击：发生于模型推理阶段，攻击者在原始输入中添加扰动，诱使已训练完成的模型输出错误结果。在BCI系统中，典型场景包括用户执行某一特定运动想象任务时，攻击者注入扰动使系统将其误识为其他意图，从而影响脑控设备的正常运行。

污染攻击：发生于模型训练阶段，攻击者在训练数据中注入带有特定模式的恶意样本，使模型在学习过程中内化后门逻辑。在系统部署后，这些逻辑可被远程触发，从而诱导模型执行非预期行为。例如，在训练EEG数据中嵌入某一特定噪声，即便该噪声与任务无关，只要数据中加入该噪声，模型仍可能输出攻击者预设的类别。

按攻击者对模型内部信息的掌握程度，对抗攻击还可分为以下两类。

白盒攻击：攻击者可完全访问目标模型的结构、参数及梯度信息，从而直接基于模型内部机制精确计算扰动方向与幅度，生成高效的对抗样本。

黑盒攻击：攻击者无法访问模型内部信息，仅能观察输入与输出。此时攻击者常通过询问黑盒模型的输入输出来估计模型的内部结构，从而构建对抗样本。

此外，根据攻击目标，对抗攻击可进一步划分为以下两类。

目标攻击：攻击者设定具体的目标标签，试图强制模型将输入样本分类为该标签。例如，在脑控字母拼写系统中，将用户的意图“字母A”误导为“字母Z”。

非目标攻击：攻击者并不指定具体输出标签，仅需令模型输出任何错误结果即可。例如，当用户执行左手运动想象时，模型被误识别为右手运动或非任务状态。

在BCI场景中，由于EEG信号本身存在低信噪比、高时变性及个体间差异性，导致BCI系统鲁棒性相对较弱，进而成为对抗攻击的高风险靶标。

对抗攻击的分类总结如表1所示。

表1 对抗攻击分类

Tab.1 Categories of Adversarial Attacks

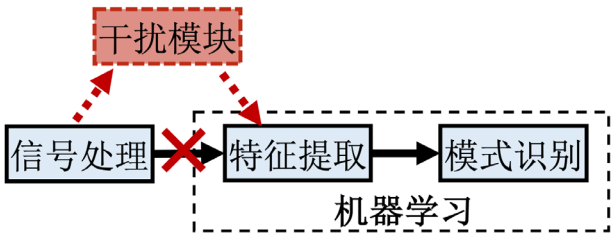
分类依据	攻击类型	特点
攻击阶段	逃逸攻击	推理阶段扰动输入，欺骗模型输出错误结果
	污染攻击	训练阶段注入污染样本，控制模型行为
攻击者知识	白盒攻击	完全知晓模型参数，精准生成对抗样本
	黑盒攻击	仅观察输入输出，通过查询推测模型行为
攻击目标	目标攻击	强制模型输出指定错误类别
	非目标攻击	仅需导致任意错误输出

1.2 逃逸攻击

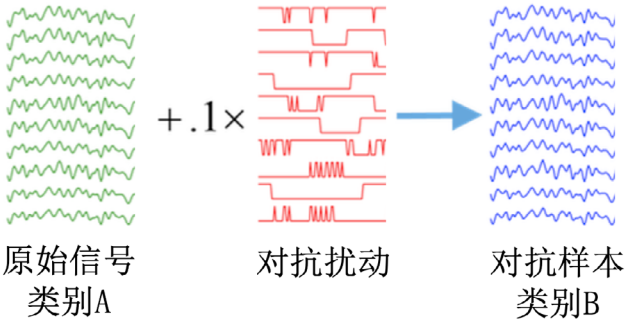
Zhang等[26]首次指出BCI中存在对抗攻击的安全隐患，并提出了一种基于无监督快速梯度符号方法的逃逸攻击框架。该方法通过在信号预处理模块与机器学习模型之间插入干扰模块，向预处理后的EEG信号注入微小扰动，从而诱导模型产生错误分类，如图2所示。针对不同的攻击知识场景，作者分别设计了白盒、灰盒和黑盒三种策略。其中，白盒攻击直接利用目标模型的损失函数梯度生成扰动；灰盒与黑盒攻击则通过构建替代模型模拟目标模型行为，并基于替代模型生成对抗样本。实验在P300、ERN和MI三种 EEG 数据集上验证了该攻击对多种卷积神经网络模型的有效性，显著降低模型分类准确率至接近随机猜测水平。该工作首次系统地揭示了BCI场景中卷积神经网络模型在对抗攻击下的脆弱性，为后续安全防护研究提供了理论基础。在此基础上，Jiang等[27]针对黑盒攻击中替代模型训练效率低的问题，提出了一种结合主动学习的二阶段攻击策略。该方法首先通过二分搜索定位靠近目标模型决策边界的样本对，随后采用正交合成策略在局部邻域内生成高信息量查询样本，从而显著减少查询次数，提高攻击效率。

Liu 等[28]提出一种基于总损失最小的通用对抗扰动生成方法，通过优化生成一个通用扰动模板，将其直接叠加到任意EEG输入上即可引发分类错误，避免了传统方法中需要对每个输入样本单独优化扰动的操作，极大提升了攻击效率与实用性。

Zhang等[29]进一步指出BCI中的传统机器学习模型同样存在对抗攻击风险。针对P300和 SSVEP 拼写器，作者分别提出了两种定制化攻击策略。对P300拼写器，通过计算非目标EEG样本在分类器损失函数上的梯度方向，构建通用扰动模板；而对SSVEP拼写器，则基于其频域特性，优化设计了周期性扰动模板，使得扰动后的信号在典型相关分析中与攻击者预设频率呈现最大相关性，从而误导模型输出。



干扰模块植入



添加对抗扰动完成攻击的过程

图2 逃逸攻击示意图

Fig.2 Schematic Diagram of Evasion Attack

Huang等[30]从频域角度切入，充分挖掘运动想象EEG信号的频谱特性，提出一种频域扰动生成方法。该方法首先将EEG信号变换至频域，在保持共轭对称性约束的前提下，针对关键频点设计微扰，并结合通道级联策略逐步优化攻击效果，从而有效欺骗卷积神经网络分类器。

除分类任务外，Meng等[31]首次探讨了 BCI 中回归模型的对抗安全问题。该工作定义了BCI回归问题中的对抗攻击范式，并将两种分类攻击策略扩展至回归场景：一种基于优化目标，另一种基于梯度信息。在驾驶疲劳监测和反应时间预测的BCI回归应用中，这两种方法均可有效操控模型输出，实现对回归值的精准偏移，揭示了回归模型在安全性方面的潜在脆弱性。

此外，上述攻击方法依赖于在信号预处理与模型之间插入干扰模块进行扰动注入，Meng等[32]进一步提出直接基于预处理滤波器实现逃逸攻击的思路。该方法通过优化生成一种通用对抗滤波器，用其对EEG信号进行滤波处理后，仅引入极小幅度的失真，便能干扰传统机器学习模型和卷积神经网络的分类结果，从而实现高效隐蔽的攻击效果。

1.3 污染攻击

Meng等[33]首次研究了BCI中的污染攻击。为了提高攻击的可实施性，该工作提出利用窄周期脉冲作为后门密钥构建污染样本，并通过污染训练数据在模型中植入隐蔽后门，从而控制训练后模型的行为。测试时，攻击者仅需在任意时刻向测试样本注入窄带周期脉冲信号即可触发误分类。在多个BCI数据集上，仅需污染5%的训练数据就能在卷积神经网络和传统模型上达到极高的攻击成功率。Jiang等[34]在此基础上进一步提出主动污染策略，通过选择最有价值的样本来构造污染样本，从而实现以最少的污染样本数量达到高效的攻击效果，极大提高了攻

击的效率和隐蔽性。

后门密钥的设计是污染攻击研究中的核心问题之一。Meng等[32]提出使用随机生成滤波器作为后门密钥。通过使用该滤波器处理少量训练样本生成污染数据，攻击者能够在模型中成功植入后门；测试阶段，任意EEG信号只需经过该滤波器处理，便可诱导模型产生预期误分类。Li等[35]从频域与空间域的角度出发，提出了一种融合频谱扰动与空间选择的后门构造策略。在频域上，利用离散傅里叶变换在特定频段内嵌入扰动，同时保持相位信息不变以提高隐蔽性；通过功率谱密度分析精确定位最优频率注入位置。在空域上，该方法基于脑区活跃性与任务相关性指标，使用动态优化算法自适应选择最优电极通道组合，实现多维扰动控制。Liu等[36]则提出一种三阶段频域污染攻击方法，系统提升后门的隐蔽性与泛化能力。第一阶段，从EEG数据中筛选易于触发攻击的样本；第二阶段，构建融合攻击成功率与隐蔽性指标的复合奖励函数，采用强化学习优化电极选择与频率扰动参数；第三阶段，基于频谱振幅插值生成污染数据，从而同时保持扰动在时域与频域的自然性，实现高质量的污染样本生成。

1.4 物理域攻击

在物理BCI系统中实施对抗攻击，通常面临诸多实际限制，如系统的时序因果性约束、扰动信号生成与注入的误差、以及攻击操作的物理可行性等。为提高攻击方法在真实系统中的可部署性，研究者围绕攻击的可实现性、鲁棒性与物理施加方式，提出了一系列针对物理域的对抗攻击策略。

针对因果性约束问题，Zhang等[29]与Liu等[28]分别提出了通用对抗扰动模板的构造方法。该类模板与具体的输入样本无关，因而适用于实时添加扰动的需求，有效提升攻击操作的实用性。

Meng等[33]提出一种基于窄带周期脉冲的后门攻击策略。该脉冲信号具备易于生成的特点，可通过电磁干扰方式直接施加于EEG采集过程中，嵌入原始信号中以触发模型误分类。此外，该方法对脉冲信号的失真具有较强的鲁棒性，即使在物理添加过程中存在噪声或失真，攻击仍具有有效性。类似地，Bian等[37]利用方波信号构造扰动，并成功攻击了基于SSVEP的脑控拼写系统，进一步验证了简单周期信号在物理攻击中的实用潜力。

Wang等[38]从优化角度出发，将物理约束直接纳入对抗扰动生成流程中。该工作基于脑电信号的空间传播模型与系统预处理步骤，构建了端到端攻击优化框架，并引入一阶导数损失项约束扰动的平稳性，使生成的对抗扰动更接近真实环境，从而提升攻击的隐蔽性与实际可添加性。

Hossen等[39]首次在真实物理系统中成功实施了对抗攻击。其方法是通过调制高频载波将模拟攻击信号耦合至EEG采集设备的模拟电路，并利用非线性电路元件完成信号解调，从而将攻击扰动叠加至原始EEG信号中。实验结果表明，该方法可显著降低系统中传统机器学习模型与卷积神经网络模型的性能，首次验证了对抗攻击在物理BCI系统中的现实威胁。Armengol-Urpi等[40]通过射频信号注入恶意脑电来操控BCI系统，成功攻击了三类EEG设备：研究级的Neuroelectrics Enobio、开源的OpenBCI Ganglion和消费级的Muse 2，演示了三种攻击场景：控制虚拟键盘输入指定字符、使无人机紧急坠毁，以及伪造深度冥想状态。该工作进一步揭示了BCI在物理域的安全漏洞，警示了未来BCI技术面临的安全挑战。

2 对抗防御

2.1 对抗防御简介

随着对抗攻击在BCI系统中所带来的安全隐患逐步受到关注，研究者提出了多种防御机制以提升系统鲁棒性。根据其核心原理与实现方式，现有对抗防御方法大致可分为以下三类：

①基于检测的防御：通过训练额外的检测器或分析模型内部激活特征，判断输入是否为对抗样本。一旦检测到潜在的异常输入，系统可采取拒绝分类、降低可信度或转入特定处理路径等策略加以应对。②基于鲁棒性训练的防御：在模型训练过程中引入对抗样本作为训练数据的一部分，使模型在学习过程中逐步适应扰动，从而提升其对恶意攻击的抵抗能力。该类方法是目前应用最广泛、效果最显著的防御策略之一。③基于输入变换的防御：通过对输入信号进行某种预处理，如滤波、投影、去噪等方式，旨在消除或削弱扰动信号在时域或频域中的表现特征，使其难以影响模型的判别结果。

Meng等[40]系统地评估了各类防御方法在BCI中的性能，构建了防御方法的性能基准体系，为后续研究提供了参考。在此基础上，Chen等[41]在欧氏对齐后的EEG数据上进一步评估了各类防御方法的有效性。

2.2 基于检测的防御

检测型防御方法试图在模型做出最终判断前识别出是否存在对抗性输入，从而防止潜在的错误决策。

Chen等[43]提出了一种基于特征距离度量的对抗样本检测方

法：该方法从神经网络最后一层提取样本特征，计算其与各类中心向量之间的马氏距离和余弦距离，并据此构建检测器。该检测器在多种白盒与黑盒攻击下均表现出高达99%的检测准确率。Zhang等[44]探索了多种不确定性指标在检测对抗样本中的作用，包括模型置信度、预测熵、贝叶斯不确定性估计与核密度估计。其中，基于核密度估计的检测器通过度量输入样本分布与训练分布的偏差，有效识别出对抗扰动，表现出良好的通用性与稳定性。此外，Khilnani等[45]引入了一种基于协方差熵的检测方法。该方法通过分析多通道EEG信号的协方差矩阵，捕捉通道间的高阶统计依赖关系，有效增强了检测模型对扰动的敏感性，从而提升了检测准确率。

2.3 基于鲁棒性训练的防御

鲁棒性训练通过在训练阶段加入对抗扰动，使模型主动适应攻击样本。

Li等[46]系统研究了五种主流对抗训练方法在BCI中的防御表现，并在三类典型白盒攻击场景下进行评估。实验表明对抗训练是提升鲁棒性的有效策略，但仍需开发更加契合EEG数据特性的防御机制。Aissa等[47]提出了一种基于生成对抗网络的对抗训练方法，首先利用快速梯度符号法生成真实对抗样本，随后通过生成对抗网络进一步生成多样化的合成对抗样本，从而提升模型对未知扰动的泛化能力。Chen等[48]将欧氏对齐与对抗训练结合，在提升BCI模型对抗鲁棒性的同时提升模型分类性能。

2.4 基于输入变换的防御

Meng等[41]根据EEG信号特性设计了一系列输入变换来削弱扰动的影响。具体地，提出了以下六种变换策略：

①随机偏移，通过左右平移信号并补零保持时序结构；②重采样，采用随机降采样后恢复原始采样率的动态处理；③通道混洗，按比例随机打乱电极通道顺序；④幅值缩放，为每个通道施加独立随机增益；⑤高斯噪声注入，按通道添加可调强度的随机噪声；⑥复合随机变换，动态组合前五种变换。

这些变换能够削弱小幅度扰动的影响，并且能够作为数据增强方式同时提升模型的泛化性能，但是对于幅度较大的扰动防御性能较弱。

3 总结及未来研究

对抗安全问题正日益成为制约BCI系统大规模实际部署的重要瓶颈。本文围绕非侵入式BCI系统中对抗攻击与防御的研究进展展开系统综述。通过梳理逃逸攻击、污染攻击和物理域攻击三个研究方向，揭示了当前BCI模型在面对不同攻击场景下的脆弱性。进一步地，本文总结了当前主流的对抗防御技术，包括基于检测的防御、基于鲁棒性训练的防御以及基于输入变换的防御。

尽管已有工作在理论与实验层面初步验证了BCI对抗攻击与防御方法的可行性，但当前研究仍存在若干亟待解决的问题。

①攻击可实施性不足：现有攻击多依赖理想化假设，尚缺乏在真实物理环境中高效施加扰动的方案。未来应加强具备时序因果性与信号可实现性的对抗方法设计，推动攻击方式向物理可操作性方向演进。②防御通用性与稳定性有限：不同防御策略在不同模型与任务中表现差异显著，缺乏可泛化至多种BCI任务的通用防御框架。未来可探索基于模型无关性与任务自适应的鲁棒性增强方法。③数据与评估标准缺失：目前尚无针对BCI对抗安全的标准化评估基准与公开测试集，难以横向比较各方法效果。构建多模态、多任务、多模型的统一评测平台将对推动领域发展起到关键作用。④攻击防御对抗博弈机制缺位：当前研究多数聚焦单一攻击或单一防御策略，尚缺乏系统性的攻防对抗建模与博弈分析。⑤结合脑认知机制的安全建模尚待深入：脑电信号具备独特的生理属性与认知结构，未来可进一步融合认知神经科学与深度学习技术，从信号源头构建更具生物可解释性的鲁棒模型，提升安全建模的科学性与合理性。

总之，BCI系统的对抗安全研究正处于快速发展阶段，其复杂性、挑战性与多学科交叉特性决定了该领域具有重要的研究价值与实践意义。未来，需在理论创新、算法设计、系统实现与应用落地等多维度协同推进，为BCI在医疗、辅助、增强等关键场景中的安全可靠运行提供坚实保障。

参考文献：

[1] VIDAL J J. Toward direct brain–computer communication[J]. Annual Review of Biophysics and Bioengineering, 1973, 2(1): 157 – 180.

[2] LORACH H, GALVEZ A, SPAGNOLO V, et al. Walking naturally after spinal cord injury using a brain–spine interface[J]. Nature, 2023, 618(7963): 126 – 133.

[3] FLESHER S N, DOWNEY J E, WEISS J M, et al. A brain–computer interface that evokes tactile sensations improves robotic arm control[J]. Science, 2021, 372(6544): 831 – 836.

[4] PARSONS B, FAUBERT J. Enhancing learning in a perceptual–

cognitive training paradigm using EEG-neurofeedback[J]. Scientific Reports, 2021, 11(1): 4061.

[5] LÜ B L, ZHANG Y Q, ZHENG W L. 情感脑机接口研究综述[J]. 智能科学与技术学报, 2021, 3(1): 36-48.

[6]《脑机接口关键科学问题、关键核心技术及其布局研究》项目组. 脑机接口技术发展现状及未来展望[J]. 科学与社会, 2024, 14(3): 2-25.

[7] LOTTE F, BOUGRAIN L, CICHOCKI A, et al. A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update[J]. Journal of Neural Engineering, 2018, 15(3): 031005.

[8] AL-SAEGH A, DAWWD S A, ABDUL-JABBAR J M. Deep learning for motor imagery EEG-based classification: A review[J]. Biomedical Signal Processing and Control, 2021, 63: 102172.

[9] LAWHERN V J, SOLON A J, WAYTOWICH N R, et al. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces[J]. Journal of Neural Engineering, 2018, 15(5): 056013.

[10] ZHANG Y, YAO S, YANG R, et al. Epileptic seizure detection based on bidirectional gated recurrent unit network[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022, 30: 135-145.

[11] ZHENG W L, LU B L. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks[J]. IEEE Transactions on Autonomous Mental Development, 2015, 7(3): 162-175.

[12] XIA K, DENG L, DUCH W, et al. Privacy-preserving domain adaptation for motor imagery-based brain-computer interfaces[J]. IEEE Transactions on Biomedical Engineering, 2022, 69(11): 3365-3376.

[13] SONG Y, ZHENG Q, LIU B, et al. EEG Conformer: Convolutional transformer for EEG decoding and visualization[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 710-719.

[14] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of the International Conference on Learning Representations. Banff, Canada: 2014: 14-16.

[15] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of the IEEE Symposium on Security and Privacy. San Jose, CA: 2017: 39-57.

[16] WEI H, TANG H, JIA X, et al. Physical adversarial attack meets computer vision: A decade survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 9797-9817.

[17] ZHOU M, WANG L, NIU Z, et al. Adversarial attack and defense in deep ranking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5306-5324.

[18] PAPERNOT N, MCDANIEL P, GOODFELLOW I. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples[EB/OL]. arXiv:1605.07277, 2016.

[19] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the International Conference on Learning Representations. San Diego, CA: 2015: 7-9.

[20] CUI X, APARCEDO A, JANG Y K, et al. On the robustness of large multimodal models against image adversarial attacks[C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Seattle, WA: 2024: 24625-24634.

[21] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[C]//Proceedings of the International Conference on Machine Learning. Scotland, UK: 2012: 1807-1814.

[22] XIAO H, BIGGIO B, BROWN G, et al. Is feature selection secure against training data poisoning?[C]//Proceedings of the International Conference on Machine Learning. Lille, France: 2015: 1689-1698.

[23] SHAFAHI A, HUANG W R, NAJIBI M, et al. Poison frogs! Targeted clean-label poisoning attacks on neural networks[C]//Proceedings of the International Conference on Neural Information Processing Systems. Montréal, Canada: 2018: 6106-6116.

[24] ALBER D A, YANG Z, ALYAKIN A, et al. Medical large language models are vulnerable to data-poisoning attacks[J]. Nature Medicine, 2025, 31: 618-626.

[25] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[EB/OL]. arXiv:1712.05526, 2017.

[26] ZHANG X, WU D. On the vulnerability of CNN classifiers in EEG-based BCIs[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2019, 27(5): 814-825.

[27] JIANG X, ZHANG X, WU D. Active learning for black-box adversarial attacks in EEG-based brain-computer interfaces[C]//Proceedings of the IEEE Symposium Series on Computational Intelligence. Xiamen, China: 2019: 361-368.

[28] LIU Z, MENG L, ZHANG X, et al. Universal adversarial perturbations for CNN classifiers in EEG-based BCIs[J]. Journal of Neural Engineering, 2021, 18(4): 0460a4.

[29] ZHANG X, WU D, DING L, et al. Tiny noise, big mistakes: Adversarial perturbations induce errors in brain-computer interface spellers[J]. National Science Review, 2020, 8(4): nwaa233.

[30] HUANG X, CHOI K S, LIANG S, et al. Frequency domain channel-wise attack to CNN classifiers in motor imagery brain-

computer interfaces[J]. IEEE Transactions on Biomedical Engineering, 2024, 71(5): 1587–1598.

[31] MENG L, LIN C T, JUNG T P, et al. White-box target attack for EEG-based BCI regression problems[C]//Proceedings of the International Conference on Neural Information Processing. Sydney, Australia: 2019: 476–488.

[32] MENG L, JIANG X, CHEN X, et al. Adversarial filtering based evasion and backdoor attacks to EEG-based brain-computer interfaces[J]. Information Fusion, 2024, 107: 102316.

[33] MENG L, JIANG X, HUANG J, et al. EEG-based brain-computer interfaces are vulnerable to backdoor attacks[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 2224–2234.

[34] JIANG X, MENG L, LI S, et al. Active poisoning: efficient backdoor attacks on transfer learning-based brain-computer interfaces[J]. Science China Information Sciences, 2023, 66(8): 182402.

[35] LI F, HUANG M, YOU W, et al. Spatial-spectral-Backdoor: Realizing backdoor attack for deep neural networks in brain–computer interface via EEG characteristics[J]. Neurocomputing, 2025, 616: 128902.

[36] LIU X H, SONG X, HE D, et al. Professor X: Manipulating EEG BCI with invisible and robust backdoor attack[EB/OL]. arXiv:2409.20158, 2024.

[37] BIAN R, MENG L, WU D. SSVEP-based brain-computer interfaces are vulnerable to square wave attacks[J]. Science China Information Sciences, 2022, 65(4): 140406.

[38] WANG X, QUINTANILLA R O S, HERSCHE M, et al. Physically-constrained adversarial attacks on brain-machine interfaces[C]//Proceedings of the Advances in Neural Information Processing Systems Workshop. New Orleans, LA: 2022.

[39] HOSSEN M I, TU Y, HEI X. A first look at the security of EEG-based systems and intelligent algorithms under physical signal injections[C]//Proceedings of the Secure and Trustworthy Deep Learning Systems Workshop. Victoria, Australia: 2023.

[40] MENG L, JIANG X, WU D. Adversarial robustness benchmark for EEG-based brain-computer interfaces[J]. Future Generation Computer Systems, 2023, 143: 231–247.

[41] CHEN X, JIA T, WU D. Data alignment based adversarial defense benchmark for EEG-based BCIs[J]. Neural Networks, 2025, 188: 107516.

[42] CHEN X, MENG L, XU Y, et al. Adversarial artifact detection in EEG-based brain-computer interfaces[J]. Journal of Neural Engineering, 2024, 21(5): 056043.

[43] ZHANG H, GU Z. Adversarial sample detection for EEG-based brain-computer interfaces[J]. Journal of Physics: Conference Series, 2024, 2761: 012037.

[44] KHILNANI A, KIRAR J S, GAUTAM G R. Enhancing adversarial attack detection in EEG signals with covariance entropy: A novel framework for BCI security[J]. IEEE Signal Processing Letters, 2025, 32: 2564-2568.

[45] LI Y, YU X, YU S, et al. Adversarial training for the adversarial robustness of EEG-based brain-computer interfaces[C]//IEEE International Workshop on Machine Learning for Signal Processing. Xi'an, China: 2022.

[46] AISSA N E H S B, KERRACHE C A, KORICHI A, et al. Enhancing EEG signal classifier robustness against adversarial attacks using a generative adversarial network approach[J]. IEEE Internet of Things Magazine, 2024, 7(3): 44-49.

[47] CHEN X, WANG Z, WU D. Alignment-based adversarial training (ABAT) for improving the robustness and accuracy of EEG-based BCIs[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2024, 32: 1703-1714.

[48] ARMENGOL-URPI A, KOVACS R, SARMA S E. Brain-Hack: Remotely injecting false brain-waves with RF to take control of a brain-computer interface[C]//Proceedings of the Workshop on CPS&IoT Security and Privacy. Copenhagen, Denmark: 2023.