Evaluating Robustness and Detection of Adversarial Attacks in EEG-Based Brain-Computer Interfaces

Beatriz C. da Costa[1], André Riker[2], Roger Immich[3], Bruno L. Dalmazo[1]

[1] Universidade Federal do Rio Grande (FURG)

[2]Universidade Federal do Pará (UFPA)

[3]Universidade Federal do Rio Grande do Norte (UFRN)

{beatriz, dalmazo}@furg.br, ariker@ufpa.br, roger@imd.ufrn.br

Abstract. Research Context: *Brain-computer interfaces (BCI) are systems that capture brain signals through techniques such as electroencephalography (EEG), processing these signals for various applications, especially in the control of devices for people with motor limitations. Despite the benefits, there are security concerns, including adversarial and cybersecurity attacks. Due to the emergence of brain-computer interaction devices on the market, it is necessary to analyze the security of these devices.* ***Scientific and/or Practical Problem:*** *Machine learning classifiers used in BCIs are vulnerable to adversarial attacks, which can compromise accuracy, safety, and user privacy. The lack of systematic evaluation of these vulnerabilities represents a gap in current research.* ***Proposed Solution and/or Analysis:*** *This work aims to emulate and analyze adversarial attacks in classifiers of BCI devices. Our experiments used the Foolbox tool to evaluate different adversarial techniques, such as DeepFool, FGSM, PGD, and Carlini-Wagner. Our evaluation identifies the negative effects of adversarial attacks on data classification.* ***Related IS Theory:*** *Technology acceptance model; Information processing theory.* ***Research Method:*** *Experiments were conducted on the BCI Competition 2008 Graz dataset, with attacks emulated during inference. Detection mechanisms based on Random Forest, SVM, and KNN were trained and evaluated to assess the feasibility of automatic defense.* ***Summary of Results:*** *Classifier accuracy decreased sharply under attack, with success rates ranging from 75.2% to 100%. Detection models achieved 83% accuracy with Random Forest and SVM for FGSM attacks, but only 5% with KNN for DeepFool, highlighting the challenge of detecting subtle perturbations.* ***Contributions and Impact to IS area:*** *The work demonstrates the vulnerabilities of BCI classifiers, proposes an evaluation pipeline for adversarial robustness, and emphasizes the importance of integrating security assessment into BCI development. Results have direct implications for information systems dealing with sensitive biomedical data.*

1. Introdução

Interfaces cérebro-computador (BCI - *Brain Computer Interfaces*) são sistemas que coletam sinais cerebrais, processam e os usam para determinado objetivo; os mais conhecidos com propósitos médicos são: controlar cadeira de rodas, robôs, braços mecânicos e teclados. Em geral, os equipamentos que possuem BCI são desenvolvidos para pacientes com algum tipo de comprometimento motor ou de locomoção.

Para o desenvolvimento de sistemas baseados em BCI precisamos obter os sinais cerebrais do paciente ou usuário. Embora existam diferentes técnicas de aquisição, o eletroencefalograma (EEG) é a técnica com menor custo e não invasiva. Nessa técnica, eletrodos são colocados no couro cabeludo e captam impulsos elétricos. O processo do uso de sinais EEG para BCI, de forma geral, envolve: (i) inicia-se com o usuário gerando atividade cerebral, (ii) pré-processamento, a fim de que seja possível extrair recursos (*e.g.* ondas cerebrais Delta, Theta, Alpha, Beta que estão relacionadas com atividades do pensamento, atividades de atenção, foco no mundo exterior e solução de problemas concretos), (iii) a classificação dos dados coletados. E em alguns casos, ainda é possível que haja um pós-processamento que seria então retornado à interface da aplicação em formato de *feedback* ao usuário.

Atualmente, existem diversos produtos no mercado com finalidades variadas, tais como jogos, análise de produtividade e análise de sentimentos, como exemplificado por [Neurable 2024]. Outros ainda estão em fase de testes, como o implante cerebral *Neuralink*[1], que busca desenvolver uma interface cérebro-computador para ajudar pessoas com paralisia, perda de memória, perda auditiva, cegueira e outros problemas neurológicos a restaurar a autonomia. Outro dispositivo em testes é da empresa *Kernel*[2] focado em medição cerebral que será capaz de acelerar o desenvolvimento de tratamentos, melhorar os resultados dos pacientes e reduzir custos com assistência médica, ofertando dados para médicos tomarem melhores decisões. Nesse cenário, com o avanço das tecnologias como EEG e dispositivos inovadores como Neuralink e Neurable, que ampliam as possibilidades de interação direta entre o cérebro e sistemas computacionais, surgem também preocupações relacionadas à segurança. Esses dispositivos tornam-se alvos potenciais para diversos ataques cibernéticos como, por exemplo, ataques adversariais, ataques de estouro de *buffer*, ataques de *malware*, entre outros [Bernal et al. 2021, Dalmazo et al. 2017, Dalmazo et al. 2018]. A literatura tem reconhecido esse problema e, nos últimos anos, diferentes trabalhos começaram a investigar a robustez dos classificadores baseados em EEG frente a ataques adversariais. Por exemplo, Xue et al. [Jiang et al. 2019] exploram cenários de ataque do tipo caixa-preta, em que o invasor apenas observa as respostas do modelo; Jiyoung et al. [Jung et al. 2023] introduziram a Generative Perturbation Network (GPN), capaz de gerar perturbações adversariais específicas para sinais de EEG; outros estudos investigam o impacto de ataques em canais específicos de EEG [Feng et al. 2021]. Apesar desses avanços, percebe-se a necessidade de explorar sistematicamente técnicas de defesa e detecção, dado o crescimento do uso desses produtos e considerando que BCIs envolvem áreas críticas como saúde, segurança e privacidade de dados.

Neste contexto, este trabalho tem como objetivo emular e analisar ataques adversariais em classificadores de dispositivos interface cérebro-computador, bem como detectar esses ataques através de técnicas de aprendizado de máquina. Em nossos experimentos, utilizamos a ferramenta Foolbox e metodologias como Método de Sinal de Gradiente Rápido (FGSM), Descida de Gradiente Projetada (PGD), DeepFool e Carlini & Wagner para introduzir ataques com intensidade controlada, examinando as respostas dos classificadores nessas condições adversas. Observamos que, em média, a acurácia do modelo

[1]https://neuralink.com/

[2]https://www.kernel.com/

foi de 76,04% antes do ataque para 25,69% após o ataque adversarial, representando uma queda significativa de 50,35%. Além disso, avaliamos três detectores de ataques adversariais baseados em: *Random Forest*, *Support Vector Machine* e *K-Nearest Neighbors* que apresentaram desempenho satisfatório para ataques FGSM moderados. No entanto, a detecção de ataques do tipo DeepFool se mostrou mais desafiadora, com desempenho consideravelmente inferior, evidenciando a complexidade em identificar perturbações mais sutis. Esses resultados demonstram não apenas a vulnerabilidade dos sistemas BCI a ataques adversariais, mas também a relevância de investigar métodos de defesa e aprimorar mecanismos automáticos de detecção para garantir a robustez e a segurança desses sistemas diante de possíveis ameaças.

O restante desse trabalho foi dividido conforme descrito. A Seção 2 apresenta os trabalhos relacionados. A arquitetura proposta é apresentada na Seção 3. A implementação, cenário de testes e resultados, na Seção 4. Considerações finais e futuros desdobramentos desse trabalho são apresentados na Seção 6.

2. Trabalhos relacionados

Neste trabalho, apresentamos uma fundamentação teórica e revisão da literatura com o objetivo de entender os estudos prévios sobre ataques adversariais em Interfaces cérebro-computador.

2.1. Fundamentação teórica: Ataques adversariais

O objetivo desse tipo de ataque é afetar a qualidade de um serviço consumido por um usuário legítimo, ou seja, deve ter um impacto leve a ponto de escapar da detecção de sistemas defensivos. Os primeiros exemplos tinham como alvo sistemas como: "controladores de admissão, servidores web e protocolos como TCP". [Kanhere and Naveed 2005]. Os avanços em sistemas de aprendizado de máquina permitiram o aperfeiçoamento dos sistemas de detecção de intrusão em diversos sistemas; no entanto, verificou-se que os algoritmos de aprendizado de máquina também poderiam ser alvos de ataques por um adversário malicioso, daí o termo "adversarial" [Barreno et al. 2006]. Um exemplo é o trabalho [Yu et al. 2010], onde os autores exploraram os impactos de ataques adversariais em sistemas de classificação que utilizavam o algoritmo K-Nearest Neighbors (KNN) e aplicaram randomização para dificultar esses ataques.

Uma área em que ataques adversariais ganharam bastante espaço foi visão computacional, como podemos constatar em [Wang et al. 2014]. Este trabalho demonstrou ataques adversariais densos e esparsos. Utilizaram um conjunto de dados de dígitos escritos à mão, onde, utilizando o dígito "7", transformaram a imagem por meio de ataque adversarial denso, modificando muitos píxeis, enquanto o ataque adversarial esparso, modificando apenas um pixel, atingiu a mesma classificação incorreta, levando o classificador a rotular como "9". Pesquisas recentes também têm analisado o impacto de ataques adversariais em outros domínios críticos, como a detecção de *malware*, demonstrando que tais ataques podem comprometer severamente a acurácia de detectores baseados em aprendizado de máquina, mesmo na presença de defesas como o treinamento adversário [Silva et al. 2024]. De acordo com [Shi et al. 2017] existem algumas maneiras de enganar um classificador por meio de ataques adversariais:

- Ataques de envenenamento (poisoning attacks): que utilizam dados de treinamento manipulados maliciosamente e introduzem no modelo legítimo.

- Ataques exploratórios: que ocorrem após o treinamento e o atacante tentam descobrir informações sobre seu funcionamento, na busca de fraquezas, como: limite de decisão, conjunto de regras, informações dos dados que foram usados no treinamento do algoritmo.
- Ataques de evasão: informam dados de teste que resultarão em uma saída incorreta.

Tratando-se do tema interfaces cérebro-computador, é possível aplicar todos os tipos de ataques adversariais, porém os principais são envenenamento e evasão, em que, no primeiro caso, um atacante pode comprometer o modelo durante o treinamento, injetando dados adversariais (manipulados) que reduzem a precisão dos classificadores. Sendo uma falha crítica para o sistema BCI que aprende com o uso a longo prazo pelo usuário seus padrões mentais. O segundo caso ocorre na fase de inferência, em que o atacante manipula as entradas que serão classificadas sem necessariamente modificar o modelo; esses ataques, no cenário de interfaces cérebro-computador, podem alterar diagnósticos médicos e controle de próteses, por exemplo. E é bastante grave para produtos que funcionam em tempo real. [Liu et al. 2018] Assim, o estudo de ataques adversariais em BCIs é essencial, pois revela vulnerabilidades críticas em sistemas que lidam diretamente com saúde, bem-estar e dados sensíveis dos usuários.

2.2. Modelo de ameaça e superfícies de ataque

O nível de conhecimento que o atacante possui sobre o sistema de aprendizado de máquina também é importante considerar, conforme descrito pelo NIST [Vassilev et al. 2024], existem três categorias principais no contexto de sistemas de aprendizado de máquina:

- **White-box**: o adversário possui conhecimento completo sobre o modelo, incluindo dados de treinamento, arquitetura e hiperparâmetros. Esse cenário representa o pior caso e é utilizado para avaliar a vulnerabilidade máxima do sistema, inclusive quando o atacante conhece eventuais mecanismos de defesa.
- **Black-box**: o adversário tem conhecimento mínimo, geralmente apenas acesso de consulta (API) ao modelo, sem detalhes sobre como foi treinado. São ataques considerados mais práticos, pois exploram interfaces de uso normal do sistema.
- **Gray-box**: representam situações intermediárias, em que o atacante conhece parcialmente o modelo (por exemplo, a arquitetura mas não os pesos, ou parte dos dados de treino). Esse cenário é bastante realista em aplicações de saúde, onde certas informações sobre o pré-processamento e os sinais utilizados podem ser de conhecimento público ou deduzidas por meio de engenharia reversa.

No contexto de BCIs, esses modelos de ameaça se traduzem em diferentes superfícies de ataque: (i) na camada física, com ruído injetado nos sensores; (ii) na transmissão, por manipulação de pacotes de dados; (iii) no armazenamento e treinamento, com envenenamento de datasets; e (iv) na inferência, por manipulação de entradas ou exploração de APIs. A definição explícita do modelo de ameaça adotado é essencial para avaliar a real viabilidade de ataques, bem como para orientar o desenvolvimento de defesas adequadas. Sendo assim, o atacante poderia treinar um modelo local com dados de banco de dados disponíveis na literatura e calcular gradientes. Com acesso de black-box, o adversário pode estimar gradientes ou ainda utilizar um modelo funcional equivalente

ao alvo a partir de respostas da API, reduzindo a lacuna para ataques. No caso do atacante possuir acesso parcial (gray-box) o atacante combina conhecimento estrutural com dados coletados no dispositivo para aproximar o gradiente efetivo. Portanto, ataques adversários no contexto de BCIs são plausíveis.

2.3. Metodologia

Fizemos uso da plataforma IEEE Xplore e ArXiv para a busca de artigos relevantes que permitissem o entendimento do estado da arte nesse tema. O IEEE Xplore é uma fonte tradicional e bem estabelecida de informações científicas, e, por se tratar de um tema recente, buscamos complementar a revisão com pesquisas mais atuais. Nesse sentido, incluímos trabalhos disponíveis no ArXiv e ACM que oferecem uma visão de estudos emergentes.

As seguintes chaves de pesquisa foram usadas nessa revisão: "Brain-Computer Interfaces", "EEG", "Adversarial attacks" e NOT "image classification" e NOT "speech recognition". As Tabelas 1 e 2 apresentam os critérios de seleção e compilam seus resultados.

Tabela 1. Critérios de inclusão (IC) e exclusão (EC)

Critério de Inclusão	
IC1	Período de Publicação entre 2019-2024
IC2	Ter acesso aberto
Critério de Exclusão	
EC1	Ser um survey
EC2	Não abordar ataques adversariais

Tabela 2. Estudos selecionados

Resultados	**Total**
IEEE Xplore	20
ArXiv	13 (dois foram retirados pois foram encontrados no IEEE Xplore)
ACM	59 (um foi retirado pois estava duplicado)
Não incluídos por IC1, IC2	1
Excluidos por EC1, EC2	60
Artigos selecionados	**31**

2.4. Análise dos Trabalhos

Entre os principais destaques, trabalhos como o de Xue, Xiao e Dongrui [Jiang et al. 2019] exploram ataques do tipo caixa preta, onde o invasor só observa as respostas do modelo às entradas, enquanto Jiyoung, HeeJoon, Geunhyeok e Hyoseok [Jung et al. 2023] propõem um modelo generativo que chamaram de GPN - *Generative perturbation network*, capaz de gerar ataques adversariais em sinais de EEG. Alguns estudos focam no impacto de ataques sobre canais específicos de EEG [Feng et al. 2021] ou na utilização de redes neurais adversariais para aumentar a robustez dos classificadores [Aissa et al. 2024]. Também existem propostas para detecção e mitigação, como

abordagens baseadas em CNNs [Aissa et al. 2023] e métodos estatísticos para identificar artefatos adversariais [Chen et al. 2024].

Bibek e Vahid [Upadhayay and Behzadan 2023] abordaram vulnerabilidades específicas de interface cérebro-computador, explorando a viabilidade de manipular especificamente BCIs de imagens motoras (MI) por meio de perturbações nos estímulos visuais e com o uso de ataques adversariais para confundir o aprendizado de máquina presente nesses sistemas. Para isso, desenvolveram um estudo próprio com sete pessoas a fim de validar a hipótese de que é possível desenvolver ataques adversariais que afetam sistemas BCI de dados de Eletroencefalografia (EEG) do tipo imagens motoras (MI).

Yunhuan, Xi , Shujian e Badong [Li et al. 2022] neste artigo os autores buscaram avaliar cinco diferentes métodos de defesas de ataques adversariais do tipo caixa branca em três paradigmas de EEG. Eles observaram que alguns métodos ainda não estão generalizados para EEG. Além disso, afirmam que os modelos ShallowCNN e DeepCNN são os mais seguros em comparação com a EEGNet. Por último, TLM-UAP é o ataque mais fraco em comparação com UFGSM e SAP.

Hoseen et al. [Hossen et al. 2023] apresentou um estudo sobre ataques de sinais eletromagnéticos modulados (EMI) capazes de enganar modelos de aprendizado de máquina, especificamente seus testes focaram em detecção de crises epiléticas e perceberam que os ataques induziram a falsos positivos. Ainda, os autores discutiram possíveis métodos de mitigação.

2.5. Discussão dos Trabalhos

Apesar dos avanços, percebe-se que a maioria das pesquisas foca em propor novos tipos de ataques adversariais e alguns poucos em desenvolver métodos de defesa. Os estudos revisados, como os de [Jung et al. 2023], que introduziram uma rede de perturbação generativa para dados de EEG, e [Feng et al. 2021], que têm como alvo canais específicos de EEG, ressaltam a necessidade de uma maior exploração dos impactos do ataque. Ao revisar os estudos selecionados, fica evidente que a maioria das pesquisas é orientada para a proposição de novos tipos de ataques adversariais. Observa-se uma lacuna significativa na exploração sistemática da eficácia de diferentes métodos de ataques adversariais e, principalmente, na investigação de técnicas de detecção. Apenas alguns estudos examinam o impacto das perturbações em vários paradigmas de EEG ou propõem abordagens práticas para identificar sinais adversariais de forma automática, aspecto crucial para a segurança de aplicações reais.

Considerando as áreas críticas frequentemente associadas aos sistemas de BCI, como saúde, segurança e tratamento de dados sensíveis, o desenvolvimento de soluções mais seguras e confiáveis torna-se fundamental, incluindo mecanismos eficazes de detecção de ataques adversariais. Portanto, análises comparativas que considerem tanto a eficácia dos ataques quanto as capacidades de defesa e detecção dos modelos são essenciais. Neste contexto, o presente trabalho busca contribuir avaliando o impacto de diferentes ataques adversariais e propondo um pipeline de detecção automática desses ataques em sinais de EEG, preenchendo uma lacuna relevante identificada na literatura.

3. Estratégia de Avaliação de Robustez em BCI

O objetivo deste trabalho é emular e analisar ataques adversariais na comunicação entre os pacientes e o BCI. A emulação do ataque é por meio de ruídos na transmissão dos sinais entre os pacientes e o sistema BCI. Além disso, propõe-se avaliar os efeitos desses ataques no desempenho dos classificadores por meio da injeção de dados falsos (conforme descrito pela Relação 1 na Figura 1) [Leite et al. 2024]. Adicionalmente, este trabalho analisa mecanismos de detecção dos ataques adversariais clássicos no contexto de interfaces cérebro-computador. Com esses objetivos em mente, os experimentos foram conduzidos utilizando um banco de dados, do qual foi extraída uma classificação, seguida da inserção dos ataques adversariais. Após os ataques, é exigida uma nova classificação (conforme Relação 2 na Figura 1), e os resultados obtidos são analisados e discutidos, buscando identificar padrões e os impactos dos ataques - bem como a eficácia do mecanismo de detecção - no ambiente de interface cérebro-computador.

Figura 1. Modelo conceitual

3.1. Dataset utilizado

O dataset utilizado é o BCI Competition 2008 Graz data set A [Brunner et al. 2024], que possui dados para diversos tipos de análises: são 9 participantes, totalizando 48 sessões. Neste trabalho foram executadas quatro tarefas de imagens motoras: mão esquerda (classe 1), mão direita (classe 2), ambos os pés (classe 3) e língua (classe 4). Para as gravações, os participantes ficavam sentados em cadeiras reclináveis, usando capacete com os eletrodos capazes de captar as informações, usando o sistema EEG-10-20 que é um padrão com até 38 canais; neste trabalho foram usados 22 desses canais, como podemos visualizar na Figura 2. Nela podemos observar a distribuição dos canais utilizados neste estudo; cada ponto representa um canal específico colocado na superfície do couro cabeludo dos participantes, com o objetivo de captar os sinais elétricos gerados pela atividade neural. Além disso, esse dataset busca estimar a influência do movimento do músculo dos olhos nos dados, por isso a sessão inicia com o participante mantendo por dois minutos os olhos

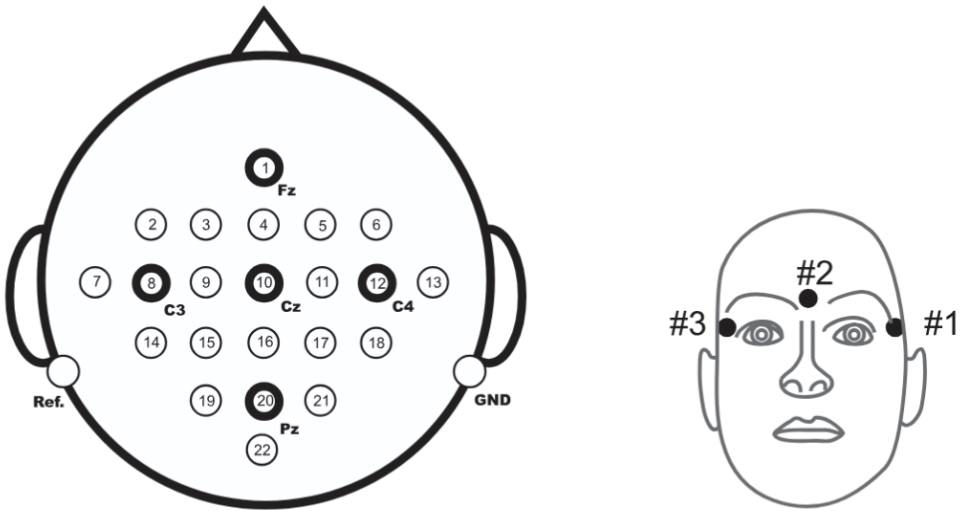

Figura 2. Canais usados para gravação dos dados

Fonte: Inspirado em [Brunner et al. 2024]

abertos e fixando o olhar em uma cruz na tela; depois, manter um minuto os olhos fechados e, por último, um minuto com movimentos oculares. Os dados foram disponibilizados em arquivos .gdf.

4. Método de pesquisa

Nesta seção, apresentamos o processo de desenvolvimento e análise de um modelo de IA que classifica um conjunto de dados de EEG, além de explorar o desempenho do classificador frente a ataques adversariais e a capacidade de detecção desses ataques.

Pré-processamento: Os dados utilizados foram o [Brunner et al. 2024], para a utilização em um classificador, primeiro é necessário a remoção de ruído e ajuste das frequências. Utilizamos a biblioteca MNE do Python, que permite a análise de dados neurofisiológicos como o EEG e outros sinais biométricos. O primeiro passo foi retirar as informações relacionadas ao EOG (*electrooculography*), pois não eram relevantes ao propósito deste trabalho, uma vez que o foco deste estudo é a análise de sinais cerebrais relacionados a atividades motoras para posteriores testes de ataques adversariais. O EOG representa atividades musculares oculares, como movimentos de olhos e piscadas, que podem introduzir artefatos indesejados, impactando negativamente o desempenho do classificador. Portanto, a remoção garante dados mais específicos ao propósito deste trabalho. Depois, aplicou-se um filtro FIR, com corte entre 7Hz e 35Hz, a fim de eliminar ruídos de baixa e alta frequência. Ainda, reduziu-se a amostragem para 200Hz, para redução de custos computacionais. Para remoção de sinais indesejados, utilizamos o Independent Component Analysis (ICA), que permite retirar ruídos de origem não cerebral. Esses sinais podem incluir atividades musculares, como contrações musculares faciais, ou também interferência elétrica de equipamentos que podem distorcer os dados do EEG. Por último, os dados foram segmentados em épocas de acordo com os eventos de imagens motoras, facilitando a análise de padrões ao longo do tempo.

Modelo: O modelo utilizado para classificação foi o EEGNET [Lawhern et al. 2018], que é uma rede neural convolucional compacta criada especificamente para aplicações de interfaces cérebro-computador. Além disso, esse modelo foi escolhido por estar presente em 14 dos 21 artigos da nossa revisão de literatura. Os hiperparâmetros utilizados também foram definidos de acordo com as referências encontradas na literatura; o treinamento empregou uma taxa de aprendizado de 0.001 e *batch size* de 64 a fim de evitar problemas relacionados ao gradiente. Na otimização, foi adotado o Adam, pois é amplamente utilizado em aprendizado profundo.

4.1. Emulação de Ataques Adversariais

Para avaliar a robustez do modelo, aplicamos técnicas de ataque adversarial aplicáveis a interfaces cérebro-computador identificadas durante a revisão da literatura, implementadas usando a biblioteca Foolbox [Rauber et al. 2020]:

- Método de Sinal de Gradiente Rápido (FGSM) [Goodfellow et al. 2015]: Esta técnica utiliza o gradiente para identificar a direção em que o valor da perda aumenta e, em seguida, aplica essa perturbação com intensidade controlada. Neste trabalho, definimos a intensidade em 0,03, um valor que permite que os dados gerados adversarialmente permaneçam dentro de uma faixa esperada dos dados originais, induzindo erros de classificação e mantendo alguma interpretabilidade dos sinais.
- DeepFool [Moosavi-Dezfooli et al. 2016]: Projetado para enganar redes neurais profundas, introduzindo perturbações mínimas nas entradas. Ele busca a menor perturbação necessária para alterar a classificação, funcionando tanto para classificadores binários quanto multiclasse.
- Gradiente Descendente Projetado (PGD) [Madry et al. 2019]: Foca em redes neurais profundas enganosas, sendo eficaz e simples de implementar. O objetivo é encontrar uma perturbação adversarial dentro de uma região restrita.
- Carlini e Wagner [Carlini and Wagner 2017]: Considerado um método altamente poderoso, ele formula a geração de exemplos adversos como um problema de minimização, suportando diversas métricas para calcular a distância:
    - Minimização das mudanças na amplitude do sinal em todas as amostras;
    - Minimização do número de pontos alterados no sinal;
    - Minimização da maior mudança em qualquer ponto do sinal.

4.2. Detecção dos Ataques Adversariais

Este trabalho avalia mecanismos de detecção desses ataques, parte central para aplicações práticas, uma vez que permite identificar tentativas de manipulação maliciosa que podem comprometer decisões sensíveis. O processo de detecção consistiu na criação de um *dataset* composto por sinais legítimos e sinais adversariais gerados via ataques FGSM e DeepFool. Os sinais passaram por técnicas de normalização e redução de dimensionalidade (como PCA) para tornar o aprendizado mais eficiente e robusto. Como estratégia de detecção, foram treinados três diferentes classificadores: Random Forest, Support Vector Machine (SVM) e K-Nearest Neighbors (KNN).

Para o Randon Forest foram utilizados 200 árvores com profundidade máxima de 15; o SVM foi configurado com kernel RBF e probabilidade ativada; e o KNN adotou 5 vizinhos com distância Euclidiana. A divisão do *dataset* para detecção foi realizada em

80% para treinamento e 20% para teste. Após o treinamento, cada detector foi avaliado no conjunto de teste de forma independente quanto à sua capacidade de distinguir amostras limpas de adulteradas. Para fins de análise detalhada, foi escolhido o ataque FGSM com $\epsilon = 0.025$, valor que demonstrou gerar degradação significativa no classificador alvo, mas ainda preservava estrutura suficiente para aprender padrões discriminativos. Nesse cenário, os três classificadores apresentaram desempenhos contrastantes: o Random Forest atingindo a maior acurácia (0.8337), seguido pelo SVM (0.8313), enquanto o KNN apresentou desempenho próximo ao aleatório (0.5), revelando uma limitação quanto à sua capacidade de generalização para ataques adversariais no contexto de EEG. As métricas completas para o FGSM estão apresentadas na Tabela 4.

Em paralelo, avaliou-se a detecção de exemplos adversariais do ataque DeepFool, e mostrou-se consideravelmente mais desafiadora, com as acurácias caindo substancialmente; o Random Forest e o SVM obtiveram menos de 13% de acurácia. O KNN atingiu acurácia levemente superior a 50%, o que ainda indica desempenho insatisfatório para distinguir sinais legítimos de adversariais. Esses resultados reforçam a dificuldade em identificar ataques com maior nível de sofisticação. A análise de variância (ANOVA) indicou diferenças estatisticamente significativas entre os classificadores ($F = 31.72$, $p < 0.0001$), embora todos tenham falhado em detectar ataques adversariais de forma eficaz, o que está alinhado com a característica do DeepFool - introduzir perturbações quase imperceptíveis nos dados.

5. Resumo dos resultados

Nesta seção, apresentamos os resultados do impacto dos ataques no classificador e o desempenho da detecção.

5.1. Impacto dos Ataques no Classificador

A Figura 3 ilustra a comparação entre o sinal de EEG original e o sinal perturbado adversamente após o ataque FGSM, destacando as perturbações sutis, porém efetivas. Vale ressaltar que a detecção desse tipo de ataque é desafiadora, pois os sinais após o ataque se assemelham bastante ao sinal original sem quaisquer perturbações.

Avaliação:Os resultados obtidos para os ataques adversários estão detalhados na Tabela 3. Comparativamente, o classificador inicial apresentou uma precisão média de 66,67% entre todos os sujeitos antes da aplicação dos ataques. As classes “mão esquerda” e “língua” foram notavelmente precisas em vários sujeitos. A implementação de diferentes técnicas adversas, como *DeepFool*, *FGSM*, *PGD* e *Carlini Wagner*, demonstrou eficácias variadas, refletidas tanto na taxa de sucesso quanto nas perturbações médias causadas por cada ataque, conforme mostrado na tabela. Para mensurar o impacto desses ataques adversários no modelo, utilizamos o método de redução percentual relativa para quantificar a queda na precisão, que avalia a proporção da diferença na precisão antes e depois do ataque em relação ao valor inicial. Por exemplo, para o Sujeito 1, a precisão antes do ataque FGSM era de 76,04% e, após o ataque, caiu para 25,69%. Aplicando a fórmula da Equação 1, isso resulta em uma redução significativa de 66,23% na precisão, indicando que o ataque adversário comprometeu gravemente o modelo ao induzir previsões incorretas generalizadas.

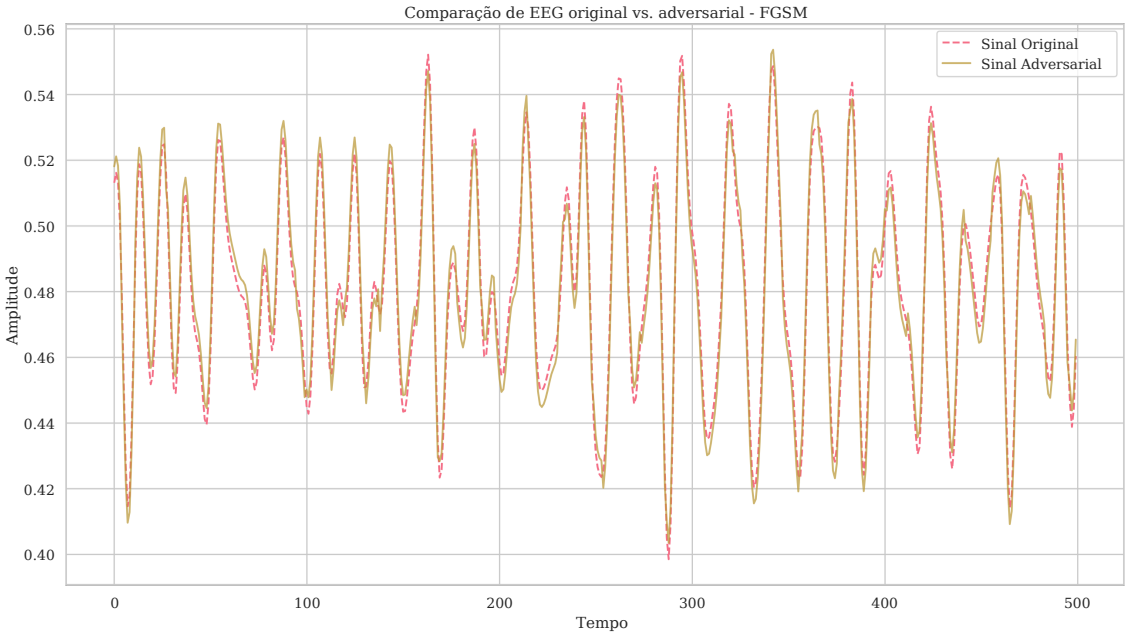

Figura 3. Comparação do sinal de EEG original e do sinal adversário pós-ataque FGSM, ilustrando as perturbações induzidas.

$$x = \frac{\text{Precisão antes do ataque} - \text{Precisão após o ataque}}{\text{Precisão antes do ataque}} \times 100 \tag{1}$$

Tabela 3. Comparação dos resultados dos ataques adversários em média para todos os sujeitos

Ataque	Precisão	Taxa de sucesso	Perturbação média
DeepFool	0.00%	100.0%	0.009303
FGSM	24.80%	75.2%	0.005000
PGD	11.4%	88.6%	0.005097
CarliniWagner	0.00%	100.0%	0.017773

Fonte: Elaborado pela autora.

- **DeepFool** e **Carlini Wagner** foram particularmente devastadores, reduzindo a precisão para 0%, indicando uma eficácia de ataque de 100% com perturbações médias de 0,009303 e 0,017773, respectivamente.
- **FGSM** e **PGD**, embora não eliminassem completamente a capacidade preditiva do modelo, ainda reduziram significativamente sua precisão para 24,8% e 11,4%, com uma taxa de sucesso de ataque de 75,2% e 88,6%, respectivamente.

Os resultados obtidos para o Sujeito 2 são ilustrados nas Figuras 4 e 5, que mostram as matrizes de confusão comparando o desempenho do modelo em dados limpos e após a aplicação de ataques adversários. Essas matrizes reforçam os resultados da tabela, mostrando o comportamento de classificação para um único sujeito e em diferentes classes. Essas observações são essenciais para entender a robustez do modelo contra ata-

ques mais sofisticados e ajudam a estabelecer uma comparação clara sobre qual técnica adversária pode ser mais prejudicial em contextos de EEG aplicados.

5.2. Desempenho da Detecção

Para avaliar como os detectores performaram, foram utilizadas diversas métricas bastante conhecidas na área de aprendizado de máquina: matriz de confusão, acurácia (*accuracy*), precisão (*precision*), *recall*, F1-score e curva ROC (*Receiver Operating Characteristic*).

As **matrizes de confusão** (Figuras 6, 7) resumem as predições dos modelos, mostrando a quantidade de acertos e erros para cada classe em cada classificador. A diagonal principal representa os acertos, enquanto os demais valores indicam erros de classificação. A análise dessas matrizes permite comparar o desempenho dos classificadores diante de exemplos limpos e adversariais.

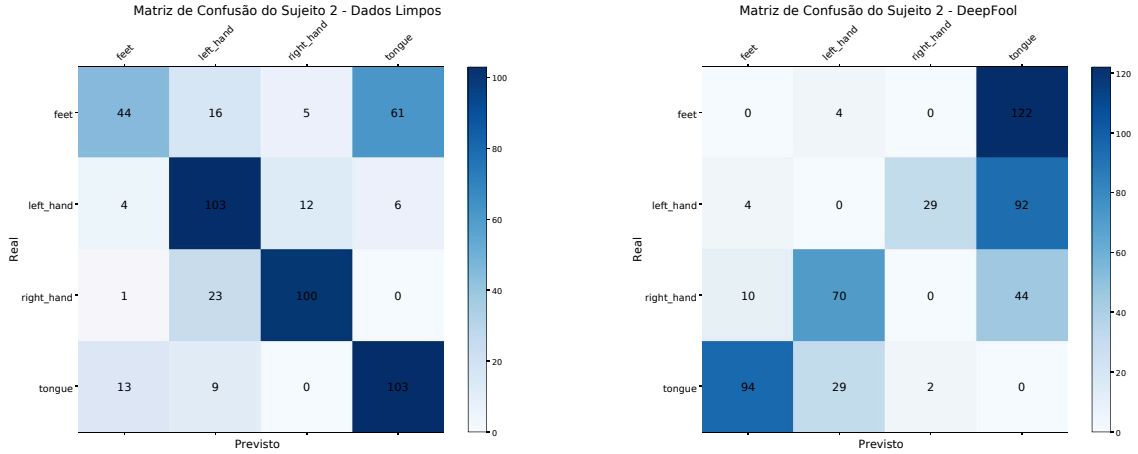

Figura 4. Matrizes de confusão para os dados limpos e para o ataque DeepFool.

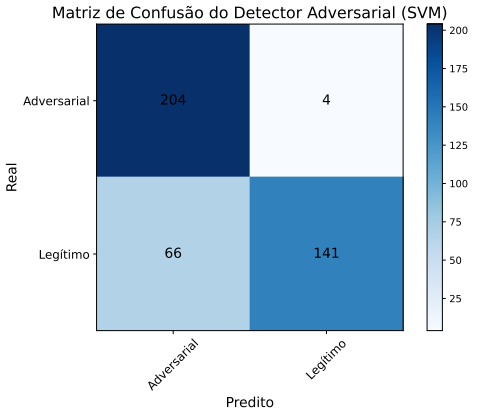

Figura 6. Matriz de confusão do classificador SVM na detecção de ataques adversariais.

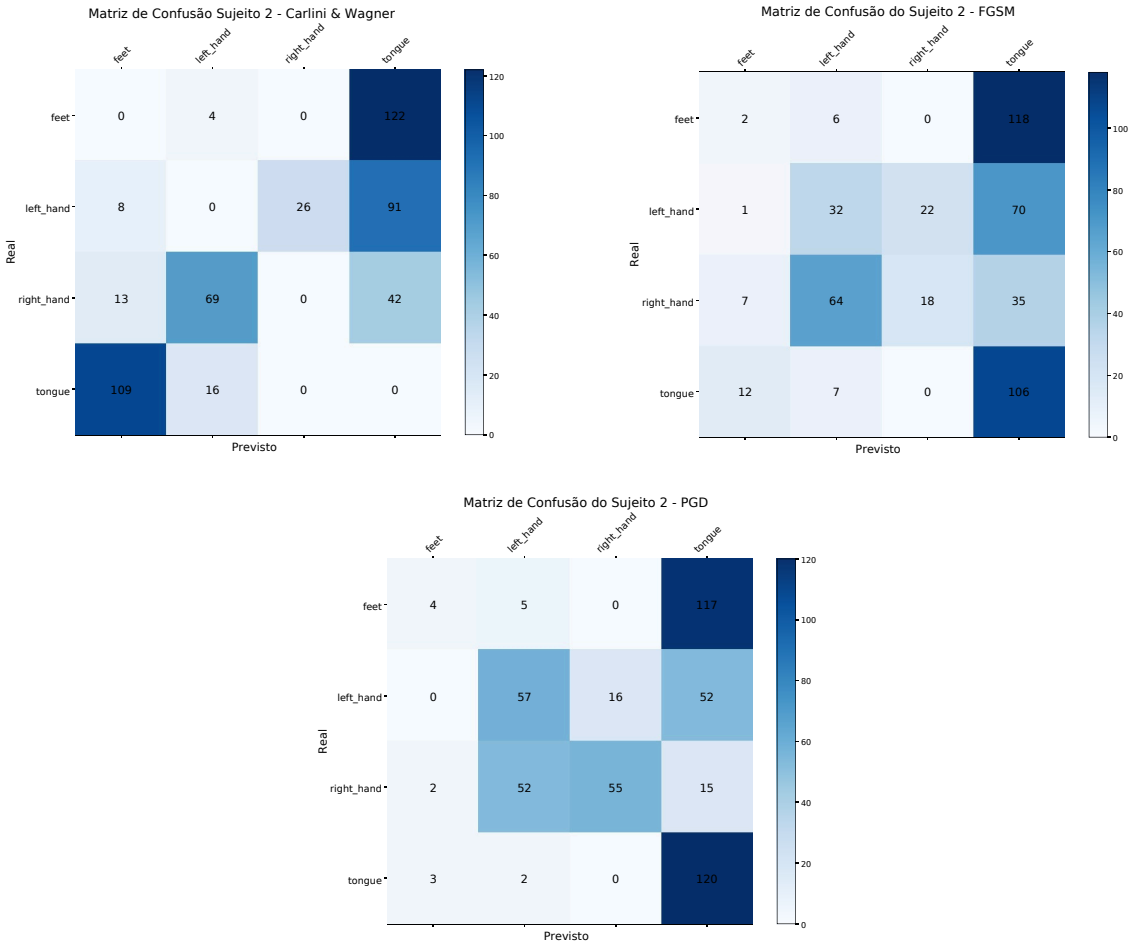

Figura 5. Matrizes de confusão para os ataques CEW, FGSM e PGD.

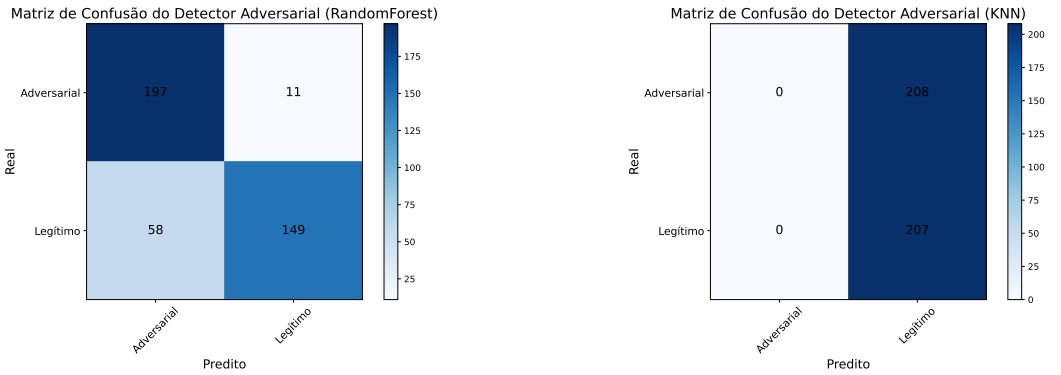

Figura 7. Matrizes de confusão para os classificadores Random Forest e KNN na detecção de ataques adversariais.

A partir desses valores, são calculadas as seguintes métricas:

- **Acurácia**: proporção de acertos totais do modelo.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

- **Precisão**: entre os exemplos classificados como positivos, quantos realmente são positivos.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- **Recall**: entre os exemplos realmente positivos, quantos foram corretamente identificados.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

- **F1-score**: média harmônica entre precisão e recall, útil para medir o equilíbrio entre as duas métricas.

$$F1\text{-}score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

Essas métricas ajudam a entender não só o quanto o modelo acerta, mas também sua capacidade de distinguir corretamente entre classes positivas e negativas. A Tabela 4 apresenta os principais indicadores de desempenho do detector para os diferentes tipos de ataques avaliados.

Tabela 4. Desempenho dos classificadores na detecção de ataques adversariais (FGSM com $\epsilon = 0.025$)

Classificador	Acurácia	Precisão	Recall	F1-score
RandomForest	0.8337	0.9313	0.7198	0.8120
SVM	0.8313	0.9724	0.6812	0.8011
KNN	0.4988	0.4988	1.0000	0.6656

Fonte: Elaborado pela autora.

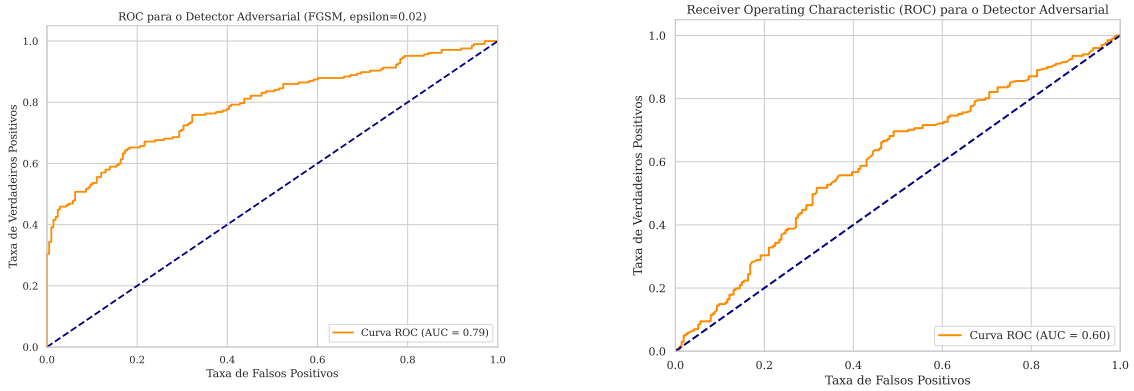

Figura 8. Curva ROC do detector Random Forest para FGSM e DeepFool.

O KNN apresentou acurácia próxima de 50%, com recall de 100%, indicando que classificou todos os exemplos como positivos, compatível com a sensibilidade do método à escala e à reordenação local de vizinhos provocada pelas perturbações pequenas. Em análises adicionais, avaliamos utilizar outros valores de k inicial como 3 ou 7 para observar o comportamento do método. Os experimentos evidenciam que, para ataques FGSM

de moderada intensidade, o mecanismo de detecção baseado em aprendizado de máquina apresenta resultados consistentes, com acurácia acima de 83% nos classificadores Random Forest e SVM, evidenciando a capacidade de distinguir amostras legítimas e adversariais de forma confiável sob ataques FGSM. Entretanto, ataques mais sutis, como o DeepFool, permanecem um desafio, com taxas elevadas de falsos negativos e positivos, o que limita a eficiência do detector nesses cenários, evidenciando a necessidade de explorar detectores mais sofisticados.

6. Considerações Finais

Neste trabalho, revisitamos a literatura sobre ataques adversariais em interfaces cérebro-computador (BCI) e realizamos experimentos para analisar os impactos desses ataques nos classificadores de sinais EEG. Utilizamos dados de EEG do banco de dados [Brunner et al. 2024], aplicando técnicas de pré-processamento e testando o classificador [Lawhern et al. 2018] tanto em condições limpas quanto após a injeção dos ataques.

Nossa avaliação de diferentes métodos de ataque — incluindo FGSM, PGD, DeepFool e Carlini-Wagner — demonstrou que mesmo perturbações mínimas podem reduzir drasticamente a acurácia dos modelos, com taxas de sucesso variando entre 75,2% e 100%. Além disso, introduzimos um pipeline de detecção para identificar a presença de ataques adversariais: geramos exemplos adulterados via DeepFool e FGSM e treinamos classificadores tradicionais como Random Forest, SVM e KNN. Entre os classificadores avaliados, para ataques gerados a partir da técnica FGSM, o Random Forest apresentou o melhor desempenho com 83,4% de acurácia, seguido de perto pelo SVM com 83,1%. Ambos atingiram acurácia superior a 83% na detecção de exemplos gerados pelo ataque FGSM, demonstrando boa capacidade de distinção entre sinais limpos e sinais adversariais, embora ainda enfrentem limitações na detecção de perturbações mais sutis,como por exemplo ataques gerados pela técnica DeepFool. O classificador KNN, por outro lado, teve desempenho significativamente inferior, evidenciando sensibilidade às perturbações, tendo sua aplicabilidade limitada no contexto de detecção de perturbações adversariais em sinais EEG. A análise foi complementada por curvas ROC e métricas adequadas ao contexto de aprendizado de máquina.

A relevância desse estudo está na crescente disponibilidade de dispositivos BCI em áreas como saúde, controle de dispositivos, jogos e foco. Nesses novos contextos, ataques adversariais podem comprometer a integridade dos dados, causando insegurança dos usuários por meio de erros em diagnósticos médicos, objetivos terapêuticos ou até mesmo frustrar a experiência do usuário com atividades de lazer, como jogos.

Nossos resultados sugerem que, embora seja viável identificar alguns ataques, ainda há necessidade de aprimorar tanto os métodos de geração de exemplos adversariais (para testar a robustez dos modelos) quanto as estratégias de detecção (por exemplo, extração de *features* tempo-frequência ou uso de *autoencoders* e CNNs). Pesquisas futuras devem explorar arquiteturas de detector mais especializadas, técnicas de redução de dimensionalidade (PCA) e outras abordagens estatísticas e de aprendizado de máquina para elevar a eficiência e a confiabilidade dos sistemas BCI frente a ameaças adversariais. Com o objetivo de assegurar a reprodutibilidade e a transparência dos resultados apresentados, disponibilizamos todos os códigos utilizados neste estudo em um repositório público no GitHub, acessível em: https://github.com/BeahIF/SBSSI2026.

Referências

Aissa, N. E. H. S. B., Kerrache, C. A., Korichi, A., Lakas, A., and Belkacem, A. N. (2024). Enhancing eeg signal classifier robustness against adversarial attacks using a generative adversarial network approach. *IEEE Internet of Things Magazine*, 7(3):44–49.

Aissa, N. E. H. S. B., Lakas, A., Korichi, A., Kerrache, C. A., and Belkacem, A. N. (2023). Robust detection of adversarial attacks for eeg-based motor imagery classification using hierarchical deep learning. In *2023 15th International Conference on Innovations in Information Technology (IIT)*, pages 156–161.

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., and Tygar, J. D. (2006). Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, ASIACCS '06, page 16–25, New York, NY, USA. Association for Computing Machinery.

Bernal, S. L., Celdrán, A. H., Pérez, G. M., Barros, M. T., and Balasubramaniam, S. (2021). Security in brain-computer interfaces: State-of-the-art, opportunities, and future challenges. *ACM Comput. Surv.*, 54(1).

Brunner, C., Leeb, R., and Müller-Putz, G. (2024). Bci competition 2008–graz data set a.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57.

Chen, X., Meng, L., Xu, Y., and Wu, D. (2024). Adversarial artifact detection in eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 21(5):056043.

Dalmazo, B. L., Vilela, J. P., and Curado, M. (2017). Performance analysis of network traffic predictors in the cloud. *Journal of Network and Systems Management*, 25:290–320.

Dalmazo, B. L., Vilela, J. P., and Curado, M. (2018). Triple-similarity mechanism for alarm management in the cloud. *Computers & Security*, 78:33–42.

Feng, B., Wang, Y., and Ding, Y. (2021). Saga: Sparse adversarial attack on eeg-based brain computer interface. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 975–979.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples.

Hossen, M. I., Tu, Y., and Hei, X. (2023). A first look at the security of eeg-based systems and intelligent algorithms under physical signal injections. In *Proceedings of the 2023 Secure and Trustworthy Deep Learning Systems Workshop*, SecTL '23, New York, NY, USA. Association for Computing Machinery.

Jiang, X., Zhang, X., and Wu, D. (2019). Active learning for black-box adversarial attacks in eeg-based brain-computer interfaces. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 361–368.

Jung, J., Moon, H., Yu, G., and Hwang, H. (2023). Generative perturbation network for universal adversarial attacks on brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 27(11):5622–5633.

Kanhere, S. and Naveed, A. (2005). A novel tuneable low-intensity adversarial attack. In *The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05)l*, pages 8 pp.–801.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013.

Leite, L., Santo, Y., Dalmazo, B., and Riker, A. (2024). Federated learning under attack: Improving gradient inversion for batch of images. In *Anais do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 794–800, Porto Alegre, RS, Brasil. SBC.

Li, Y., Yu, X., Yu, S., and Chen, B. (2022). Adversarial training for the adversarial robustness of eeg-based brain-computer interfaces. In *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., and Leung, V. C. M. (2018). A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE Access*, 6:12103–12117.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582.

Neurable (2024). Neurable. Company website.

Rauber, J., Zimmermann, R., Bethge, M., and Brendel, W. (2020). Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 5(53):2607.

Shi, Y., Sagduyu, Y., and Grushin, A. (2017). How to steal a machine learning classifier with deep learning. In *2017 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–5.

Silva, G., Junior, G. F., and Zarpelão, B. (2024). Impacto de ataques de evasão e eficácia da defesa baseada em treinamento adversário em detectores de malware. In *Anais do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 829–835, Porto Alegre, RS, Brasil. SBC.

Upadhayay, B. and Behzadan, V. (2023). Adversarial stimuli: Attacking brain-computer interfaces via perturbed sensory events. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3061–3066.

Vassilev, A., Oprea, A., Fordyce, A., and Anderson, H. (2024). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. NIST Artificial Intelligence (AI) Report NIST AI 100-2e2023, National Institute of Standards and Technology, Gaithersburg, MD.

Wang, F., Liu, W., and Chawla, S. (2014). On sparse feature attacks in adversarial learning. In *2014 IEEE International Conference on Data Mining*, pages 1013–1018.

Yu, H., Chan, P. P. K., Ng, W. W. Y., and Yeung, D. S. (2010). Apply randomization in knn to make the adversary harder to attack the classifier. In *2010 International Conference on Machine Learning and Cybernetics*, volume 1, pages 179–183.