XXI Escola Regional de Redes de Computadores / VIII Workshop de Segurança (ERRC). DOI: 10.5753/errc.2024.4583

ARTIGO COMPLETO/FULL PAPER

Avaliando Ataques Adversariais em Aplicações de Interfaces Cérebro-Computador

Evaluating Adversarial Attacks in Applications of Brain-Computer Interfaces

Beatriz Costa • beatrizdacosta1@furg.br
Universidade Federal do Rio Grande (FURG)

André Riker • ariker@ufpa.br
Universidade Federal do Pará (UFPA)

Bruno L. Dalmazo • dalmazo@furg.br
Universidade Federal do Rio Grande (FURG)

RESUMO. Interfaces cérebro-computador (BCI- do inglês brain computer interface) são sistemas que captam sinais cerebrais através de técnicas como eletroencefalografia (EEG), processando esses sinais para diversas aplicações, especialmente no controle de dispositivos para pessoas com limitações motoras. Apesar dos benefícios, há preocupações com segurança, incluindo ataques adversarias e de segurança cibernética. A proteção dos dados pessoais dos usuários é crucial, exigindo pesquisas contínuas em cibersegurança para mitigar riscos e garantir a privacidade nas aplicações BCI. Durante a avaliação, podemos identificar efeitos negativos na classificação de dados produzidos por ataques adversariais. Em virtude do surgimento de dispositivos de interação cérebro computador no mercado, que acessam ondas cerebrais dos usuários e os analisam para diversos propósitos, se faz necessário a análise do quesito segurança desses dispositivos. Diante disto, este trabalho tem como objetivo emular e analisar ataques adversariais em classificadores de dispositivos interface cérebro computador.

ABSTRACT. 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 adversary and cybersecurity attacks. Protecting users' personal data is crucial, requiring continuous research into cybersecurity to mitigate risks and ensure privacy in BCI applications. During the evaluation, we can identify negative effects on data classification produced by adversarial attacks. Due to the emergence of brain-computer interaction devices on the market, which accesses users' brain waves and analyzes them for various purposes, it is necessary to analyze the security of these devices. Therefore, this work aims to emulate and analyze adversarial attacks in classifiers of brain-computer interface devices

PALAVRAS-CHAVE: Segurança • interface cérebro computador • ataques adversariais
KEYWORDS: Security • brain-computer interface • adversarial attacks

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, robô, braço mecânico, teclado, desenvolvidos para pacientes com algum tipo de comprometimento motor ou de locomoção.

Para o desenvolvimento desses sistemas precisamos dos sinais cerebrais do paciente ou usuário, algumas técnicas para essa coleta existem, mas eletroencefalograma (EEG) é a técnica com menos custo e não invasiva. Essa técnica utiliza eletrodos colocados no couro cabeludo que captam impulsos elétricos. O processo do uso de sinais EEG para BCI, de forma geral, inicia-se com o usuário gerando atividade cerebral, depois esses dados passam por um 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 e foco no mundo exterior e solução de problemas concretos), que depois serão classificados. Dependendo do tipo de aplicação, ainda é possível que haja um pós-processamento que seria então retornado a interface da aplicação em formato de feedback ao usuário.

Nesse cenários, alguns ataques de cibersegurança podem ocorrer, como por exemplo, ataques adversariais, ataques de estouro de buffer, ataques de malware, entre outros [1–3]. Atualmente, existem diversos produtos de interface cérebro computador disponíveis no mercado com finalidades variadas, tais como jogos, análise de produtividade e análise de sentimentos, como

exemplificado por [4]. Outros ainda estão em fase de testes, como o implante cerebral *Neuralink* [5], 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, o sistema de medição cerebral será capaz de acelerar o desenvolvimento de tratamentos, melhorar os resultados dos pacientes e reduzir custos com assistência médica, ofertando dados para médicos terem melhores decisões e ainda o sistema *Neurosity* que será capaz de analisar o foco e controlar interfaces digitais a partir da mente em tempo real [6].

Dado o crescimento do uso desses produtos e o processo em que os dados são transmitidos dos dispositivos EEG para plataformas de análise, como computadores ou até mesmo a nuvem, onde são processados e armazenados, torna-se imperativo discutir e avaliar aspectos de segurança dessas aplicações. A transmissão e armazenamento de informações pessoais dos usuários envolve riscos significativos à privacidade e segurança dos usuários e dados. Portanto, é de vital importância que a comunidade científica investigue e desenvolva medidas robustas para proteger esses dados, garantindo que as inovações tecnológicas não comprometam a privacidade e segurança dos usuários. Neste contexto, esse trabalho tem como objetivo emular ataques *adversariais* na comunicação entre os usuário e o sistema BCI e analisar seus efeitos nos classificadores.

2 Trabalhos Relacionados

Neste trabalho foi desenvolvida uma revisão da literatura com o objetivo de entender os estudos prévios sobre ataques adversariais em aplicações cérebro computador.

2.1 Metodologia

Fizemos uso da plataforma IEEE Xplore, com as seguintes chaves de pesquisa: “Brain-Computer Interfaces” e “Adversarial attacks”. As Tabelas 1 e 2 apresentam os critérios de seleção e compilam seus resultados, enquanto a Tabela 3 apresenta os artigos resultantes.

A Tabela 3 apresenta os artigos resultantes e nessa seção discute-se brevemente os artigos selecionados.

Xue, Xiao e Dongrui [7] focaram em construir um ataque de caixa preta onde o invasor só pode observar as respostas do modelo às entradas, diferente de outros trabalhos anteriores onde o atacante possui informações sobre a arquitetura, parâmetros e dados de treinamento do modelo.

Jiyoung, HeeJoon, Geunhyeok e Hyoseok [8] propõem um modelo generativo que chamaram de GPN - *Generative perturbation network*, capaz de gerar ataques adversariais. Esse modelo é capaz de produzir perturbações gerais e também específicas para EEG, sendo os primeiros que aplicam essa abordagem com foco em EEG.

Bibek e Vahid [9] 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 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 (EG) do tipo imagens motoras (MI).

Boyuan, Yuke e Yufei [10] realizaram ataques adversariais esparsos, que chamaram de SAGA (*Sparse Adversarial eeG Attack*), pois focaram no fato de que cada canal EEG representa uma determinada área do cérebro, sendo assim buscam analisar o impacto da inserção de perturbações em diferentes canais.

Nour, Chaker, Ahmed, Abderrahmane e Abdelkader [11] propõem um ataque adversarial baseado em uma GAN (*generative adversarial network*) com o propósito de aumentar a robustez de sistemas BCI, os resultados do trabalho demonstram melhorias no desempenho dos classificadores.

Yunhuan, Xi , Shujian e Badong [12] 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.

Nour , Abderrahmane, Ahmed, Chaker e Abdelkader [13] neste estudo os autores realizaram uma investigação de ataques adversariais do tipo FGSM no modelo EEGNET, propõem uma detecção desse tipo de ataque baseada em CNN com objetivo de distinguir dados limpos e dados com perturbações adversariais.

Essam, Nour e Monica [14] neste estudo os autores buscaram uma nova abordagem sobre ataques adversariais em sinais de EEG, focaram na preservação da privacidade, ou seja, usar de perturbações adversariais para alterar informações confidenciais a fim de impedir acessos não autorizados.

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

Fonte: Elaborado pela autora.

Tabela 2. Estudos selecionados

Resultados	**Total**
IEEE Xplore	20
Não incluídos por IC1, IC2	1
Excluidos por EC1, EC2	11
Artigos selecionados	**8**

Fonte: Elaborado pela autora.

2.2 Discussão dos Trabalhos

Ao analisar os artigos selecionados exibidos na Tabela 3, observa-se uma lacuna sobre estudos comparativos de ataques adversariais em diferentes classificadores para EEG. Identificar quais ataques performam melhor em determinados classificadores pode ser um objeto de interesse para a melhoria de dispositivos de BCI já disponíveis no mercado e também novos que poderão surgir, assegurando maior confiabilidade dos dados. Os estudos comparativos de ataques adversariais em diferentes classificadores para EEG podem contribuir para novas abordagens de vulnerabilidades de dispositivos BCI, que permitirão melhores defesas contra este tipo de ataque. Portanto, explorar o desempenho de ataques adversariais é essencial para promover mais confiança no uso dessas tecnologias.

3 Proposta

O objetivo deste trabalho é analisar ruídos que emulem um ataque adversarial na comunicação entre os pacientes e o BCI e analisar efeitos nos comportamentos dos classificadores com a injeção de dados falsos (descrito pela relação 1 na Figura 1) [15]. Os experimentos foram conduzidos utilizando um banco de dados. O ataque então é inserido no tráfego dos dados e uma nova classificação é exigida (relação 2 na Figura 1). Por fim, os resultados obtidos são analisados e discutidos, buscando-se identificar padrões, avaliar a eficiência em identificar esse tipo de ataque no ambiente de interface cérebro computador.

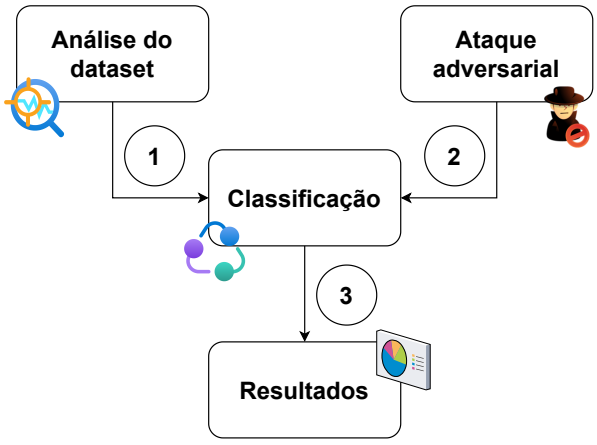


Figura 1. Modelo conceitual

Fonte: Elaborado pela autora.

3.1 Dataset utilizado

O dataset utilizado é o [16], que possui dados para vários tipo de análises: são 60 horas de gravações EEG, de 13 participantes, totalizando 75 sessões. Neste trabalho foram 60.000 exemplos de imagens motoras em quatro diferentes paradigmas. Com participantes entre 20 e 35 anos, 8 pessoas do sexo masculino e 5 do sexo feminino, saudáveis, sem condições psiquiátricas, uso de medicamentos e contraindicações ao EEG. Para as gravações os participantes ficavam sentados em cadeiras reclináveis usando capacete com os eletrodos capazes de captar as informações, usando o sistema EEG-1200 que é um padrão com até 38 canais, neste trabalho foram usados 19 desses canais como podemos visualizar na Figura 2a. Nela podemos observar a distri-

Tabela 3. Trabalhos selecionados

N	Título	Ano
01	Active Learning for Black-Box Adversarial Attacks in EEG-Based Brain-Computer Interfaces	2019
02	Generative Perturbation Network for Universal Adversarial Attacks on Brain-Computer Interfaces	2023
03	Adversarial Stimuli: Attacking Brain-Computer Interfaces via Perturbed Sensory Events	2023
04	Saga: Sparse Adversarial Attack on EEG-Based Brain Computer Interface	2021
05	Enhancing EEG Signal Classifier Robustness Against Adversarial Attacks Using a Generative Adversarial Network Approach	2024
06	Adversarial Training for the Adversarial Robustness of EEG-Based Brain-Computer Interfaces	2022
07	Robust Detection of Adversarial Attacks for EEG-based Motor Imagery Classification using Hierarchical Deep Learning	2023
08	A Privacy-Preserving Generative Adversarial Network Method for Securing EEG Brain Signals	2020

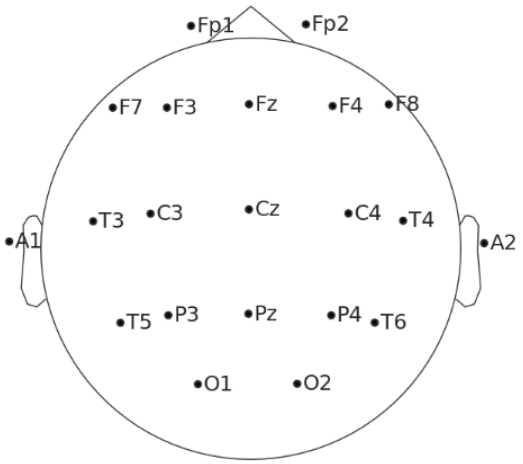
Fonte: Elaborado pela autora.

buição dos canais utilizados neste estudo, cada ponto representa um canal específico colocado na superfície do couro cabeludo dos participantes, com objetivo de captar os sinais elétricos gerados pela atividade neural. Na Figura 2b, podemos visualizar as telas que eram apresentadas aos participantes, implementada no Matlab, que mostra diferentes imagens com movimentos a serem imaginados pelos participantes. Os dados foram disponibilizados em arquivos .matlab, organizados de acordo com a Tabela 4.

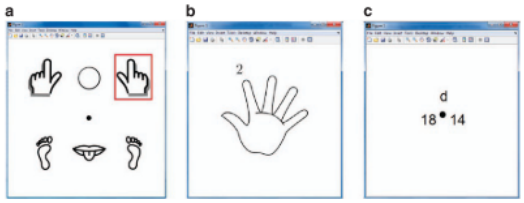
Tabela 4. Dados BCI EEG [16]

Variável	Descrição
id	Identificador alfanumérico exclusivo do registro
nS	Número de amostras de dados de EEG
sampFreq	Frequência de amostragem dos dados de EEG
marker	eGUI [registro da sessão de gravação]
data	Dados de EEG da sessão de gravação

Fonte: Elaborado pela autora.



(a) Canais usados para gravação dos dados em [16]



(b) Interface para gravação dos dados em [16]

Figura 2. Ambiente de de gravação de dados.

Fonte: [16]

4 Avaliação e Implementação

De acordo com [17], exitem diferentes frameworks para desenvolvimento de aplicações de interface cérebro computador, no trabalho [18] os autores selecionaram o MNE, pacote do Python que permite visualizar e analizar dados neurofisiológicos humanos de MEG (Magnetoencefalografia), EEG, ECoG (Eletrococleografia), NIRS (Espectroscopia Infravermelha Próximo) entre outros. Como mencionado na 1 com posse dos dados de

EEG é possível, reconhecer emoções, atenção, doenças e até mesmo intenção de movimento. O banco de dados escolhido é focado em Imagens Motoras [19], onde o usuário imagina o movimento de uma parte do corpo sem realmente executá-lo fisicamente. O trabalho iniciou com a análise os dados do banco de dados [16], aplicou-se um pré-processamento em código Python, para enfim testar o classificador [20], que deve ser capaz de identificar padrões cerebrais que podem corresponder ao movimento de determinada parte do corpo. Depois foi gerado o ataque adversarial [8], que cria uma perturbação nos dados originais, alterando o comportamento do classificador. Essa técnica induz o modelo a erros e é importante especialmente em sistemas com dados sensíveis, como interfaces cérebro-computador. Na Figura 3 é possível visualizar os resultados, por meio de uma matriz de confusão que compara o desempenho do classificador em dois cenários: dados limpos e dados alterados (após ataque). Cada quadrante da matriz indica quantas vezes a classe foi prevista corretamente ou incorretamente pelo modelo. Observa-se, por exemplo, que mesmo a classe 2, com número alto de previsões corretas ainda sofre com erros após o ataque adversarial. Percebe-se que o ataque influencia o classificador a confundir a classe ”Left foot” com a classe ”Right foot”, com uma taxa de 67, sugerindo que o ataque adversarial, além de gerar previsões incorretas de maneira geral, também indica a possibilidade de uma classificação incorreta intencional, caso o atacante tenha essas informações previamente e tente explorar o controle do sistema.

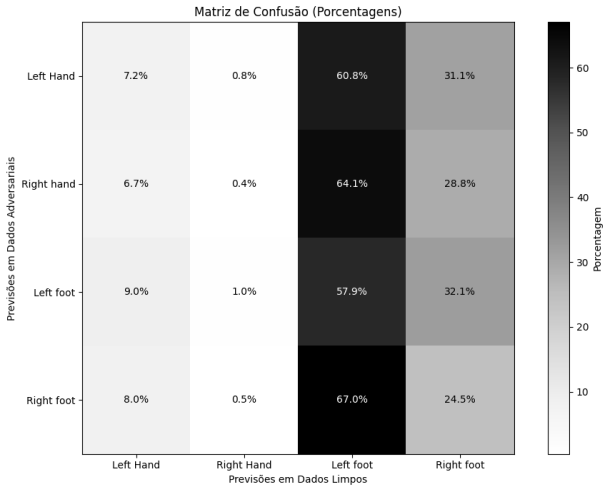


Figura 3. Matriz de confusão comparando o desempenho do classificador em dados limpos e dados pós ataque adversarial

Fonte: Elaborado pela autora.

Ainda, na Figura 4 podemos visualizar o impacto do ataque adversarial no classificador em uma classe específica. A partir dessa matriz de confusão, podemos analisar o modelo teve dificuldades de identificar a classe “hand”, com apenas alguns exemplos (429) sendo classificado corretamente.

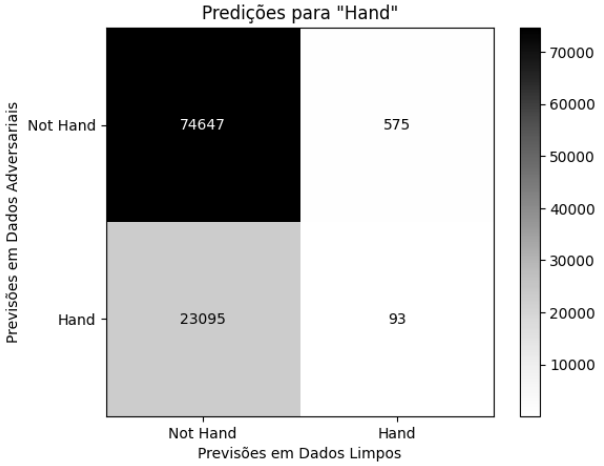


Figura 4. Matriz de confusão comparando o desempenho do classificador em dados limpos e dados pós ataque adversarial em uma classe específica

Fonte: Elaborado pela autora.

5 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 [16], aplicando técnicas de pré-processamento e testando um classificador antes e após os ataques. Os resultados revelam um impacto significativo na capacidade do modelo de distinguir as diferentes classes, o que interfere diretamente na interpretação dos sinais cerebrais pela interface e pode levar à falhas. Nesses sistemas de interface cérebro computador a precisão é essencial, pois ataque adversariais podem comprometer o funcionamento de diversos dispositivos com propósitos médicos:controle de cadeira de rodas, próteses, entre outros e também com propósitos recreativos ou de performance, mas que carregam informações sigilosas. A vulnerabilidade do modelo foi evidenciada para a classe “hand” que sofreu maior índice de erros, podendo resultar em por exemplo, um movimento incorreto de uma prótese. Esses resultados indicam a necessidade de aprimorar os classificadores, buscando uma robustez contra ataques adversariais. Para pesquisas futuras, pretendemos analisar quais características e informações extraídas dos dados EEG têm maior influência no desempenho dos

modelos, a fim de mitigar esses ataques garantindo sistemas BCI mais seguros e eficazes.

Declarações complementares

Financiamento

Os autores agradecem à FAPERGS (23/2551-0000773-8).

Referências

1 Bernal, S. L. *et al.* Security in Brain-Computer Interfaces: State-of-the-Art, Opportunities, and Future Challenges. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 54, n. 1, jan. 2021. ISSN 0360-0300. DOI: 10.1145/3427376. Disponível em: https://doi.org/10.1145/3427376.

2 Dalmazo, B. L.; Vilela, J. P.; Curado, M. Performance analysis of network traffic predictors in the cloud. *Journal of Network and Systems Management*, Springer, v. 25, p. 290–320, 2017.

3 Dalmazo, B. L.; Vilela, J. P.; Curado, M. Triple-Similarity Mechanism for alarm management in the cloud. *Computers & Security*, v. 78, p. 33–42, 2018. ISSN 0167-4048. DOI: https://doi.org/10.1016/j.cose.2018.05.016. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167404818306515.

4 Neurable. *Neurable*. 2024. Disponível em: https://www.neurable.io/.

5 Neuralink. *Neuralink*. 2024. Disponível em: https://neuralink.com/.

6 Neurosity. *Neurosity*. 2024. Disponível em: https://neurosity.co/.

7 Jiang, X.; Zhang, X.; Wu, D. Active Learning for Black-Box Adversarial Attacks in EEG-Based Brain-Computer Interfaces. *In*: 2019 IEEE Symposium Series on Computational Intelligence (SSCI). 2019. P. 361–368. DOI: 10.1109/SSCI44817.2019.9002719.

8 Jung, J. *et al.* Generative Perturbation Network for Universal Adversarial Attacks on Brain-Computer Interfaces. *IEEE Journal of Biomedical and Health Informatics*, v. 27, n. 11, p. 5622–5633, 2023. DOI: 10.1109/JBHI.2023.3303494.

9 Upadhayay, B.; Behzadan, V. Adversarial Stimuli: Attacking Brain-Computer Interfaces via Perturbed Sensory Events. *In*: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2023. P. 3061–3066. DOI: 10.1109/SMC53992.2023.10394505.

10 Feng, B.; Wang, Y.; Ding, Y. Saga: Sparse Adversarial Attack on EEG-Based Brain Computer Interface. *In*: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. P. 975–979. DOI: 10.1109/ICASSP39728.2021.9413507.

11 Aissa, N. E. H. S. B. *et al.* Enhancing EEG Signal Classifier Robustness Against Adversarial Attacks Using a Generative Adversarial Network Approach. *IEEE Internet of Things Magazine*, v. 7, n. 3, p. 44–49, 2024. DOI: 10.1109/IOTM.001.2300262.

12 Li, Y. *et al.* 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). 2022. P. 1–6. DOI: 10.1109/MLSP55214.2022.9943479.

13 Aissa, N. E. H. S. B. *et al.* 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). 2023. P. 156–161. DOI: 10.1109/IIT59782.2023.10366492.

14 Debie, E.; Moustafa, N.; Whitty, M. T. A Privacy-Preserving Generative Adversarial Network Method for Securing EEG Brain Signals. *In*: 2020 International Joint Conference on Neural Networks (IJCNN). 2020. P. 1–8. DOI: 10.1109/IJCNN48605.2020.9206683.

15 Leite, L. *et al.* 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. São José dos Campos/SP: SBC, 2024. P. 794–800. DOI: 10.5753/sbseg.2024.241680. Disponível em: https://sol.sbc.org.br/index.php/sbseg/article/view/30070.

16 Kaya, M. *et al. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces.* 2018. Wiki do abnTeX2. Disponível em: https://figshare.com/collections/A_large_electroencephalographic_motor_imagery_dataset_for_electroencephalographic_brain_computer_interfaces/3917698.

17 Beltrán, E. T. M. *et al.* SecBrain: A Framework to Detect Cyberattacks Revealing Sensitive Data in Brain-Computer Interfaces. *In*: ADVANCES in Malware and Data-Driven Network Security. IGI Global, 2022. P. 176–198.

18 Martínez Beltrán, E. T. *et al.* Noise-based cyberattacks generating fake P300 waves in brain–computer interfaces. *Cluster Computing*, Springer, p. 1–16, 2021.

19 Amorim, M. *et al.* Systematic Review of Aggregation Functions Applied to Image Edge Detection. *Axioms*, v. 12, n. 4, 2023. ISSN 2075-1680. DOI: 10.3390/axioms12040330. Disponível em: https://www.mdpi.com/2075-1680/12/4/330.

20 Lawhern, V. J. *et al.* EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, IOP Publishing, v. 15, n. 5, p. 056013, jul. 2018. DOI: 10.1088/1741-2552/aace8c. Disponível em: https://dx.doi.org/10.1088/1741-2552/aace8c.