Explorando Perturbações Adversariais em BCIs via Detectores Baseados em Sistemas Fuzzy

Beatriz C. da Costa[1], **Giancarlo Lucca**[2], **Helida Santos**[1], **Bruno L. Dalmazo**[1]

[1] Centro de Ciências Computacionais
Universidade Federal do Rio Grande (FURG)

[2]Universidade Católica de Pelotas (UCPel)

{beatrizdacosta1, helida, dalmazo}@furg.br, giancarlo.lucca@ucpel.edu.br

Abstract. *Brain-computer interface (BCI) enable direct communication between the brain and computational systems, with applications in healthcare, entertainment, and accessibility. However, EEG signal classification models are vulnerable to adversarial attacks, which introduce small perturbations into the signals to mislead the classifiers. This work investigates the impact of such attacks and the performance of detection mechanisms based on machine learning algorithms and fuzzy systems. To this end, we implemented a fuzzy system and compared it with Random Forest and SVM detectors, using FGSM and DeepFool attacks on the EEGNet classifier trained with data from the BCI Competition IV 2a. Our results show that it is possible to detect moderate attacks with good performance, although more subtle attacks remain a challenge. The experiments highlight the need for more robust detection methods in BCI environments.*

Resumo. *Interface cérebro-computador (do inglês, Brain-Computer Interface - BCI) permite a comunicação direta entre o cérebro e sistemas computacionais, com aplicações em saúde, entretenimento e acessibilidade. No entanto, modelos de classificação de sinais EEG estão sujeitos a ataques adversariais, que introduzem pequenas perturbações nos sinais para enganar os classificadores. Este trabalho investiga o impacto desses ataques e o desempenho de mecanismos de detecção baseados em algoritmos de aprendizado de máquina e sistemas fuzzy. Para isso, foi implementado um sistema fuzzy e comparado com detectores Random Forest e SVM, utilizando ataques FGSM e DeepFool sobre o classificador EEGNet treinado com dados da BCI Competition IV 2a. Os resultados deste trabalho indicam que é possível detectar ataques moderados com bom desempenho, embora ataques mais sutis ainda representem um desafio. Os experimentos evidenciam a necessidade de métodos de detecção mais robustos em ambientes BCI.*

1. Introdução

Interface cérebro-computador (BCI - *Brain Computer Interfaces*) possibilita a comunicação direta entre o cérebro humano e sistemas computacionais, com aplicações em saúde, acessibilidade e entretenimento. Dentre as técnicas de aquisição de sinais cerebrais, o eletroencefalograma (EEG) destaca-se por ser de baixo custo e não invasivo. Após a aquisição, os sinais passam por etapas de pré-processamento, extração de características e classificação, podendo ainda gerar feedbacks ao usuário [Santo et al. 2023].

Nos últimos anos, BCIs ganharam espaço em produtos comerciais além de pesquisas médicas, como exemplificado pelos dispositivos da Neurable, Neuralink e Kernel. No entanto, essa expansão traz preocupações com a segurança, dado que os sistemas BCI operam com dados sensíveis e em tempo real. Ataques adversariais - pequenas perturbações nos sinais EEG projetadas para enganar classificadores - podem comprometer diagnósticos ou o controle de próteses, por exemplo [Bernal et al. 2021, Dalmazo et al. 2017, Dalmazo et al. 2018, Antunes et al. 2022].

Neste contexto, este trabalho tem como objetivos emular e analisar ataques adversariais em classificadores de dispositivos interface cérebro-computador e propor mecanismos de detecção desses ataques. Para isso, aplicamos diferentes técnicas adversariais (FGSM, PGD, DeepFool e Carlini & Wagner) para introduzir ataques com intensidade controlada sobre o classificador EEGNet. Além disso, avaliamos três detectores de ataques adversariais baseados em: *Random Forest*, *Support Vector Machine* e *baseado em sistema fuzzy* que apresentaram desempenho satisfatório para ataques FGSM moderados. No entanto, a detecção de ataques do tipo DeepFool mostrou-se mais desafiadora, com desempenho consideravelmente inferior, evidenciando a complexidade em identificar perturbações mais sutis. O restante deste artigo 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 estão na Seção 4. Considerações finais e futuros desdobramentos deste trabalho são apresentados na Seção 5.

2. Trabalhos Relacionados

Estudos sobre ataques adversariais em sistemas de aprendizado de máquina demonstram que pequenas perturbações nas entradas são suficientes para induzir erros em classificadores, com impacto relevante em aplicações críticas, como visão computacional. Em BCIs, ataques do tipo evasão e envenenamento são especialmente preocupantes, pois podem alterar decisões clínicas ou comandos em tempo real de próteses [Liu et al. 2018]. Diversos trabalhos já demonstraram a eficácia de ataques como FGSM, PGD, DeepFool e Carlini & Wagner em modelos utilizados em BCIs, como os analisados por [Jiang et al. 2019] e [Jung et al. 2023]. Algumas abordagens propõem mecanismos de robustez, como treinamento adversarial ou redes neurais generativas [Aissa et al. 2024], enquanto outras propõem detectores baseados em redes neurais convolucionais (CNN) [Aissa et al. 2023]. Recentemente, sistemas fuzzy vêm sendo explorados como alternativas promissoras na detecção de ataques adversariais [Li et al. 2024]. Os detectores fuzzy baseiam-se em regras que permitem descrever o grau de similaridade entre sinais originais e corrompidos, sendo capazes de detectar ataques sutis com boa generalização.

3. Proposta

Este trabalho tem como objetivo analisar os impactos de ataques adversariais em classificadores de sinais EEG utilizados em interfaces cérebro-computador, bem como propor e comparar mecanismos de detecção desses ataques. Os ataques são emulados por meio da inserção de ruídos artificiais nos sinais, utilizando técnicas como FGSM e DeepFool. Avaliamos a robustez dos classificadores antes e após os ataques e a eficácia de três detectores: Random Forest, SVM e um sistema fuzzy. A Figura 1 ilustra o modelo conceitual da abordagem adotada: após o pré-processamento dos sinais e treinamento do classificador (EEGNet), ataques são aplicados aos dados e, em seguida, é realizada uma etapa de

detecção. Os resultados permitem avaliar a vulnerabilidade do sistema e a capacidade de identificar sinais comprometidos.

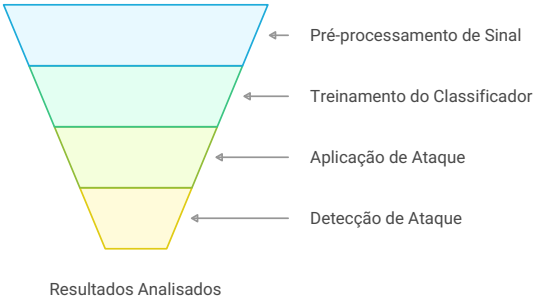

Figura 1. Modelo conceitual

4. Materiais e Métodos

Pré-processamento: Utilizamos o dataset BCI Competition IV 2a [C. Brunner and Pfurtscheller1 2008], processado com a biblioteca MNE. Sinais de eletrooculografia (EOG), que introduzem artefatos relacionados a movimentos oculares, foram removidos por não serem relevantes ao objetivo da análise motora. Aplicamos um filtro FIR entre 7-35 Hz e reduzimos a taxa de amostragem para 200 Hz. Para mitigar ruídos não neurais, empregamos Análise de Componentes Independentes (ICA). Por fim, os dados foram segmentados em épocas com base nos eventos de imaginação motora.

Classificação e ataques adversariais: O modelo EEGNET [Lawhern et al. 2018] foi utilizado como classificador. Em seguida, aplicamos ataques adversariais com a biblioteca Foolbox, incluindo FGSM, PGD, DeepFool e Carlini & Wagner. A acurácia foi comparada antes e depois dos ataques.

Avaliação: Avaliações foram conduzidas para medir o impacto de ataques adversariais no desempenho do classificador EEGNet. A acurácia média inicial entre os sujeitos foi de 66,67%. Após os ataques, observou-se redução significativa na performance.

Como ilustrado na Tabela 1, os ataques DeepFool e Carlini & Wagner foram os mais danosos, reduzindo a acurácia a 0% com perturbações médias de 0,0093 e 0,0177, respectivamente. Já os ataques FGSM e PGD reduziram a acurácia para 24,8% e 11,4%, com sucesso superior a 75%.

Tabela 1. Comparação dos resultados dos ataques adversários

Ataque	Precisão antes	Taxa de sucesso	Perturbação média
DeepFool	0%	100%	0,009303
FGSM	24,80%	75,2%	0,005000
PGD	11,4%	88,6%	0,005097
Carlini & Wagner	0%	100%	0,017773

Os resultados obtidos evidenciam a vulnerabilidade do modelo frente a diferentes ataques, sendo especialmente relevante para aplicações críticas em tempo real. As figuras 3 e 2 exemplificam o efeito desses ataques para o Sujeito 2.

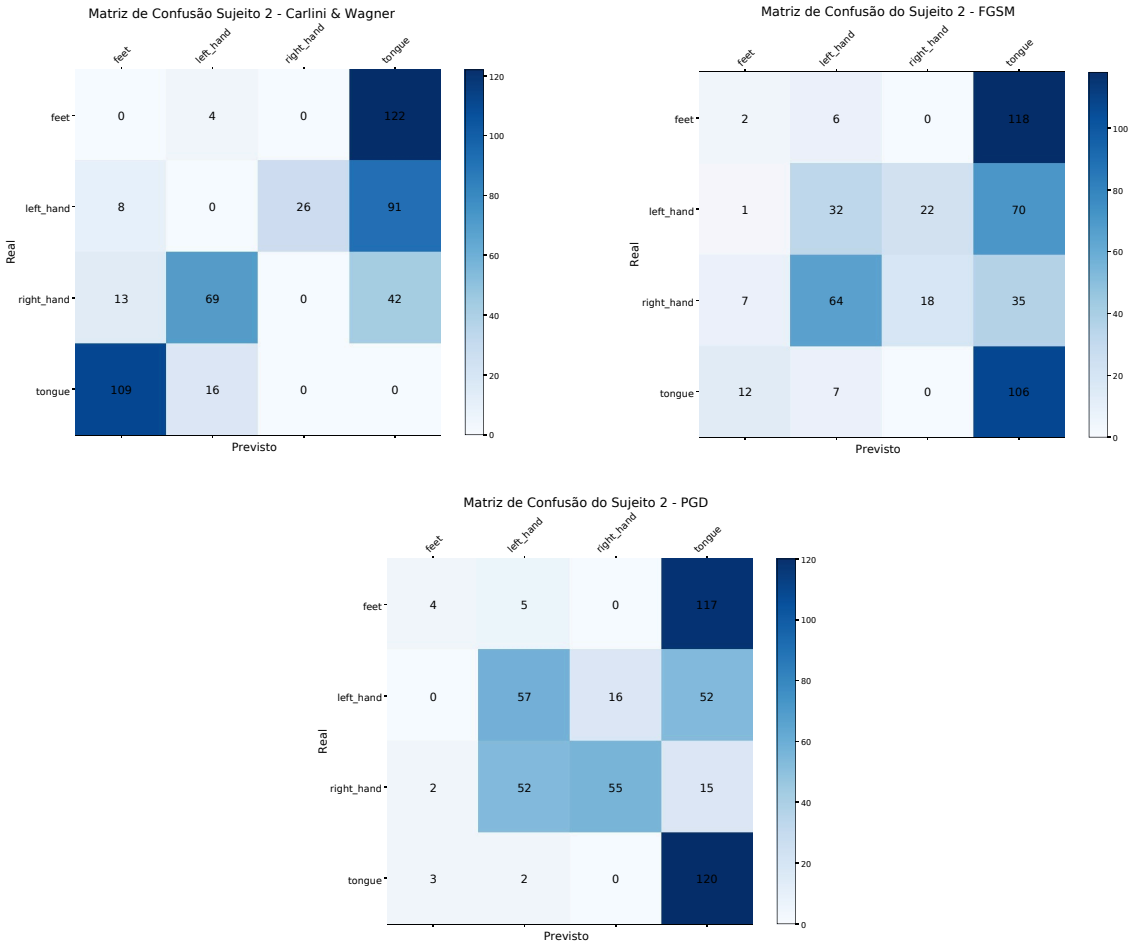

Figura 3. Matrizes de confusão para os ataques CEW, FGSM e PGD.

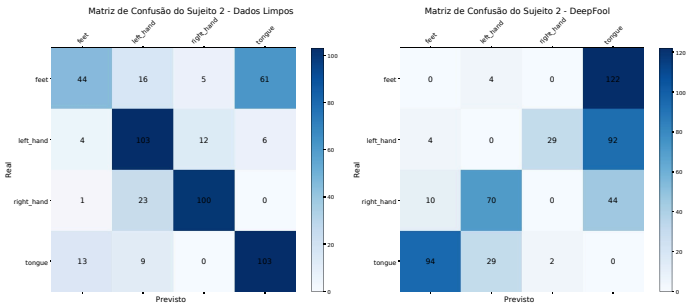

Figura 2. Matrizes de confusão para os dados limpos e para o ataque DeepFool.

4.1. Detecção dos Ataques Adversariais

Dada a vulnerabilidade dos classificadores de EEG a ataques adversariais, avaliamos diferentes mecanismos de detecção capazes de distinguir sinais legítimos de adulterados. Foi construído um dataset misto com amostras limpas e perturbadas por FGSM e DeepFool. Após normalização e redução de dimensionalidade via PCA, foram testados quatro

detectores: *Random Forest*, *SVM*, *k-Nearest Neighbors (KNN)* e um sistema fuzzy. O ataque FGSM ($\epsilon = 0{,}025$) foi utilizado como cenário principal de avaliação, por gerar perturbações suficientemente impactantes, mas ainda detectáveis. Os melhores desempenhos foram obtidos por Random Forest (acurácia de 0,83) e SVM (0,83). O KNN, por outro lado, teve desempenho inferior, com acurácia próxima a 0,5. Além dos modelos tradicionais, avaliamos um detector fuzzy simples, baseado em uma função de pertinência triangular aplicada sobre o erro quadrático médio (*Mean Squared Error - MSE*) entre amostras limpas e adversariais.

A abordagem fuzzificou os valores de MSE em quatro termos linguísticos: *very clean*, *clean*, *noisy* e *very noisy*. Em seguida, aplicou-se defuzzificação ponderada para gerar escores contínuos, classificando como ataque sinais acima de um limiar. Contudo, os resultados iniciais foram insatisfatórios: em 50 exemplos adversariais, apenas 2 foram detectados. Para aprimorar o sistema, introduziu-se uma janela deslizante de agregação, inspirada em trabalhos recentes [Ayres et al. 2024]. Cada janela foi avaliada pela média dos escores fuzzy, e a decisão passou a considerar múltiplas amostras. Com esse ajuste, a taxa de detecção aumentou de 4% para 78,26%, evidenciando que a agregação temporal eleva a sensibilidade do modelo frente a ataques sutis.

5. Considerações Finais

Este trabalho investigou os efeitos de ataques adversariais em classificadores de sinais EEG aplicados a interfaces cérebro-computador (BCI), com foco em sua detecção. Utilizamos dados do BCI Competition IV 2a e ataques como FGSM, PGD, DeepFool e Carlini & Wagner, que reduziram drasticamente a acurácia dos modelos, com taxas de sucesso superiores a 75%.

Para mitigar esse problema, avaliamos três detectores tradicionais (Random Forest, SVM e KNN) e um sistema fuzzy. Random Forest e SVM foram eficazes contra ataques moderados como FGSM (acima de 83% de acurácia), mas falharam diante de ataques mais furtivos como DeepFool. Já o sistema fuzzy, inicialmente limitado, mostrou avanços após ajustes simples, como o uso de janela deslizante, revelando potencial quando combinado a mecanismos de agregação contextual. A técnica pode ser aprimorada com integrais fuzzy (como Choquet), funções de pertinência otimizadas e regras adaptativas, indicando caminhos para uma detecção mais robusta e interpretável em BCIs.

Concluímos que ataques adversariais representam um risco real para sistemas BCI, e que estratégias de detecção interpretáveis ainda carecem de aprimoramento técnico. Trabalhos futuros devem explorar abordagens baseadas em tempo-frequência, autoencoders, CNNs e modelos fuzzy mais sofisticados - por exemplo, sistemas fuzzy intervalares para lidar com incertezas maiores e arquiteturas híbridas que combinem fuzzy com aprendizado profundo para otimizar parâmetros e regras de forma adaptativa.

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 Int. 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 classi-

fication using hierarchical deep learning. In *2023 15th International Conference on Innovations in Information Technology (IIT)*, pages 156–161.

Antunes, R. A., Dalmazo, B. L., and Drews, P. L. J. (2022). Detecting data injection attacks in ros systems using machine learning. In *2022 Latin American Robotics Symposium (LARS), 2022 Brazilian Symposium on Robotics (SBR), and 2022 Workshop on Robotics in Education (WRE)*, pages 1–6.

Ayres, D., Quevedo, A., Lucca, G., Dimuro, G., and Dalmazo, B. (2024). Comparando médias móveis com integral de choquet para detectar anomalias no tráfego de redes. In *Anais Estendidos do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 353–357, Porto Alegre, RS, Brasil. SBC.

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).

C. Brunner, R. Leeb, G. R. M.-P. A. S. and Pfurtscheller1, G. (2008). Bci competition 2008 graz data set a.

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.

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.

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.

Li, Y., Angelov, P., and Suri, N. (2024). Adversarial attack detection via fuzzy predictions. *IEEE Transactions on Fuzzy Systems*, 32(12):7015–7024.

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.

Santo, Y., Immich, R., Dalmazo, B. L., and Riker, A. (2023). Fault detection on the edge and adaptive communication for state of alert in industrial internet of things. *Sensors*, 23(7).