Adversarial Detection in EEG-Based BCIs: A Comparative Study of Classical and Neuro-Fuzzy Approaches

Beatriz Conceição da Costa[1(✉)], Giancarlo Lucca[1], Lizandro de Souza Oliveira[2], Rafael Berri[1], Roger Immich[3], Eduardo Borges[1], Richard F. Pinto[1], Fabian C. Cardoso[4], and Bruno L. Dalmazo[1]

[1] Universidade Federal do Rio Grande (FURG), Rio Grande, Brazil
{beatrizdacosta1,giancarlo.lucca,rafaelberri,eduardoborges,dalmazo}@furg.br
[2] Universidade Católica de Pelotas (UCPel), Pelotas, Brazil
lizandro.oliveira@ucpel.edu.br
[3] Universidade Federal do Rio Grande do Norte (UFRN), Natal, Brazil
roger@imd.ufrn.br
[4] University of Rio Verde (UniRV), Rio Verde, Brazil
fabian@unirv.edu.br

Abstract. Brain–Computer Interfaces (BCIs) based on electroencephalography (EEG) allow users to interact with external systems through brain activity and are increasingly used beyond clinical applications. As these systems rely on machine learning and process sensitive neural data, their resilience to malicious manipulation becomes a key concern. Adversarial attacks are especially problematic, as they introduce subtle perturbations that can significantly disrupt model predictions. This paper evaluates the impact of adversarial attacks on EEG-based BCI classifiers and investigates detection strategies related to this domain. Using the BCI Competition IV 2a dataset and EEGNet as a baseline model, we emulate FGSM, PGD, and DeepFool attacks with the Foolbox framework. Results show consistent performance degradation under all attacks, confirming the vulnerability of EEG-based classification pipelines. To address this issue, we compare traditional machine learning detectors with an Adaptive Neuro-Fuzzy Inference System (ANFIS). The results indicate that detector effectiveness is related to the attack: ANFIS performs best under FGSM and PGD, while Random Forest achieves the strongest results for DeepFool in the current setup. However, due to the limited number of effective adversarial examples and the single-subject evaluation, these findings should be interpreted as preliminary evidence rather than as a general benchmark. Even when not leading in accuracy, ANFIS remained competitive in AUC-based evaluation and provided a continuous adversarial risk score, which may support uncertainty-aware decision making in safety-critical BCI systems.

Keywords: Brain-computer interfaces · adversarial attacks · EEG · cybersecurity · neuro-fuzzy systems

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 173–184, 2027.
https://doi.org/10.1007/978-3-032-30488-9_11

1 Introduction

Advances in technology have allowed Brain-Computer Interfaces (BCIs) [20] to help people with motor disabilities control wheelchairs and prosthetic devices. These systems often use electroencephalography (EEG) [19], a non-invasive and low-cost method that records electrical activity from the scalp. The recorded signals are processed, converted into meaningful features, and analyzed by machine learning classifiers to determine the user's intentions.

Recently, BCIs have expanded beyond clinical and assistive applications, finding use in entertainment, cognitive enhancement, and emotion analysis, as demonstrated by Neurable[1]. As these systems become more widespread and share personal and sensitive data between different places, concerns over their security and trustworthiness become increasingly relevant.

These devices become potential targets for various cyberattacks [24], such as adversarial attacks [22], buffer overflow attacks [16], malware attacks [15] and others [4,10,13,17]. Among emerging threats to BCIs, adversarial machine learning poses a urgent challenge.

Adversarial attacks are manipulations crafted to deceive machine learning models by introducing smal perturbations that are often imperceptible to humans. These attacks can be categorized into poisoning, exploratory, and evasion attacks [21], with the latter two being especially relevant in BCI contexts. In poisoning attacks, adversaries inject manipulated data during training, while evasion attacks target the model during inference. Despite the growing awareness of cybersecurity in AI-driven applications, the specific vulnerabilities of BCI systems to adversarial perturbations remain underexplored. This gap highlights the need for targeted investigations into both the feasibility of attacks and the development of detection mechanisms tailored to EEG-based classification pipelines.

This work contributes an early experimental evaluation of adversarial attack impact and detection strategies in EEG-based BCI classification. Rather than claiming a definitive comparative benchmark, the study focuses on analyzing how different detectors behave under distinct adversarial perturbation regimes and on exploring the potential of neuro-fuzzy inference as an uncertainty-aware complement to classical machine learning detectors. To provide a structured analysis, the proposed approach is organized into three main components:

1. By using the Foolbox tool, we have applied attacks such as the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and DeepFool on EEG data to assess classifier robustness.
2. By analyzing three classical classifiers (Random Forest (RF) [6], Support Vector Machine (SVM) [8], and K-Nearest Neighbors (KNN) [9]).
3. By exploring the use of fuzzy logic as an uncertainty-aware detection mechanism. Specifically, an Adaptive Neuro-Fuzzy Inference System (ANFIS) [11] is employed to model adversarial risk as a continuous score rather than a binary decision.

[1] https://www.neurable.com//.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 describes the experimental setup and the proposed detection strategy. Section 4 reports the results and discussion. Finally, Sect. 5 concludes the paper.

2 Related Work

In this section, we present a literature review aimed at understanding previous studies on adversarial attacks in Brain-Computer Interfaces (BCIs), focusing on attack generation, impact analysis, and detection methods.

2.1 Search Methodology

In this study, a Systematic Literature Review was conducted to answer the following research question:

> *How are adversarial attacks applied to EEG-based brain-computer interface classifiers, and how effective are existing detection strategies in mitigating their impact on system robustness?*

For this, we used the IEEE Xplore platform to search for relevant articles that would enable an understanding of the state of the art in this topic, using the query: "Brain-Computer Interfaces AND Adversarial Attack AND detector AND fuzzy logic". IEEE Xplore is a traditional and well-established source of scientific information. We considered studies published between 2020 and 2025, excluding surveys and papers that did not address adversarial attacks.

A total of 14 papers was selected for detailed review after filtering duplicates and applying the above criteria. Table 1 compiles their results.

Table 1. Selected Studies

Results	Total
IEEE Xplore	28
Not included due to IC1, IC2	3
Excluded due to EC1, EC2	11
Selected Articles	**14**

2.2 Analysis and Discussion of Selected Works

Recent studies investigate adversarial threats to electroencephalography (EEG)-based Brain–Computer Interface (BCI) pipelines and report substantial performance degradation in commonly used deep models under small perturbations. More advanced attacks exploit structured representations, for example time–frequency features, to deceive brainprint-based systems [14]. Data poisoning has

also been explored, including label-flipping in federated learning settings, showing significant drops in model accuracy [3].

Research on defenses proposes detection architectures that combine clean-signal classification with adversarial detection [1]. Federated-learning scenarios have been studied with augmented training data, including electroencephalography samples generated by Generative Adversarial Networks (GANs) and multiple classifier backbones, reporting mixed robustness depending on the model and data regime [2]. Related efforts also leverage GAN-based anomaly detection to identify deviations in EEG patterns [5].

Fuzzy-based approaches are comparatively rare. Among the selected studies, [23] proposes a hierarchical TakagiâĂŞSugenoâĂŞKang fuzzy classifier with adversarial learning for epileptic EEG classification, indicating potential benefits of fuzzy reasoning in noisy and uncertain EEG settings.

Overall, the literature concentrates on attack crafting, adversarial training, and deep learning-based detectors, while uncertainty-aware neuro-fuzzy detection remains underexplored. In particular, the use of Adaptive Neuro-Fuzzy Inference Systems (ANFIS) have not been widely investigated for adversarial detection in EEG-based BCIs. This work addresses this gap by evaluating an ANFIS-based detector using deep-model features under strong attacks (PGD, DeepFool, and FGSM).

3 Experimental Setup

This section describes the experimental configuration adopted to evaluate adversarial robustness and detection in EEG-based BCI classification. First, we present the conceptual organization of the proposed pipeline. Next, we describe the dataset, preprocessing steps, and baseline classifier, followed by the adversarial attack generation procedure and the neuro-fuzzy detection strategy used in the experiments.

3.1 Conceptual Model

The proposed pipeline follows a structured four-steps roadmap, as illustrated in Fig. 1. (i) First, a baseline EEG classifier is established to set the performance benchmark for clean signal recognition. (ii) This is followed by a stage of adversarial emulation, where signal perturbations are generated to systematically challenge the model's robustness. (iii) Subsequently, a specialized detection system is implemented to identify these adversarial inputs in real-time. (iv) Finally, the framework concludes with a performance analysis phase that quantifies both the resulting accuracy degradation of the classifier and the overall effectiveness of the detection mechanism.

To evaluate the efficacy of the proposed framework under controlled conditions, we conducted the following experimental setup.

Dataset: We utilize the BCI Competition IV 2a dataset [7], which contains recordings from 9 subjects across four motor-imagery classes (left hand, right hand, feet, and tongue) using 22 EEG channels.

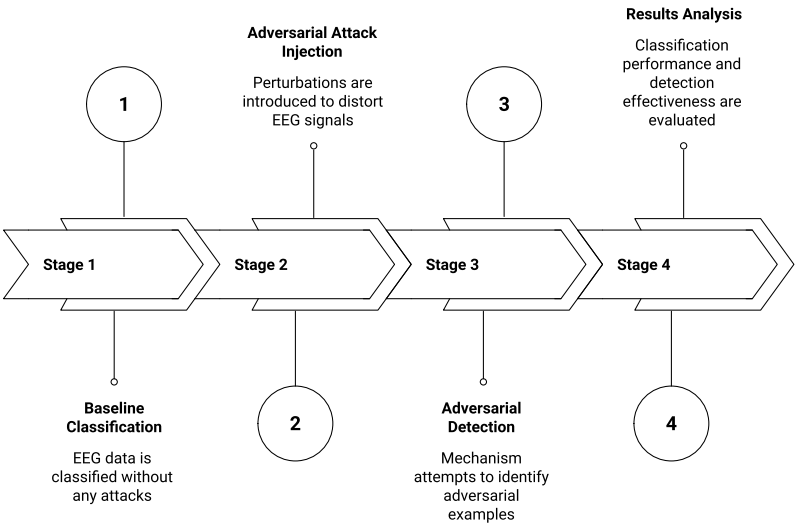

Fig. 1. Conceptual roadmap of the proposed framework

Preprocessing: The raw EEG signals are bandpass filtered (7–35 Hz) to isolate the relevant mu and beta rhythms, downsampled to 200 Hz, and cleaned using Independent Component Analysis (ICA) for artifact removal. Task-related epochs are then segmented using the MNE-Python library.

Classifier: EEGNet [12] is adopted as the baseline model because of its efficiency in extracting spatial and temporal features through depthwise and separable convolutions. Training is conducted using the Adam optimizer ($lr = 0.001$, batch size 64).

3.2 Adversarial Attack Emulation

Adversarial examples are generated using Foolbox [18]. We simulate a threat landscape by employing three standard attacks:

- **FGSM and PGD**: Gradient-based methods that introduce perturbations along the direction of the loss function's gradient.
- **DeepFool**: A geometry-based approach that calculates the minimal perturbation required to cross the decision boundary.

3.3 Adversarial Detection with Fuzzy Inference

Detection is formulated as a binary classification problem (*clean* vs. *adversarial*) using features extracted from EEGNet's output layer. While conventional detectors (SVM, RF) provide rigid boundaries, our approach utilizes an Adaptive Neuro-Fuzzy Inference System (ANFIS) [11] to explicitly represent the uncertainty inherent in perturbed BCI signals.

Table 2. Impact of adversarial attacks on EEGNet.

Attack	Clean Acc.	Adv. Acc.	Success Rate	Avg. L_2 Pert.
FGSM	0.690	0.509	0.190	4.452
PGD	0.690	0.526	0.172	4.034
DeepFool	0.690	0.509	0.181	2.601

Fuzzy models are well suited to this domain because adversarial noise is often minimal, leading to overlapping distributions where"hard" decision boundaries are unreliable. By mapping classifier outputs into linguistic concepts such as *low confidence* or *high anomaly*, ANFIS produces a continuous adversarial risk score. This is particularly relevant for safety-critical BCIs, where ambiguous cases should be flagged with calibrated confidence rather than strict binary labels.

4 Results and Discussion

This section evaluates the robustness of the EEGNet classifier under adversarial attacks and compares classical and neuro-fuzzy detectors for identifying adversarial EEG samples.

4.1 Attack Impact on EEGNet

The EEGNet baseline achieved a test accuracy of 68.97% on clean samples, providing a baseline reference point for the subsequent adversarial analysis. Under attack, performance decreased consistently across all evaluated methods, confirming the vulnerability of EEG-based BCI classification to adversarial perturbations. As shown in Table 2, FGSM and DeepFool reduced the classifier accuracy to 50.86%, while PGD resulted in 52.59%. Although the attack success rates were relatively close in this experimental setting, DeepFool produced the smallest average perturbation ($L_2 = 2.601$), suggesting that it can induce misclassification with lower overall distortion.

4.2 Detector Comparison Across Attack Families

After establishing classifier vulnerability, we evaluated four adversarial detectors: Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and ANFIS. In contrast to the previous version of this work, the current evaluation includes cross-attack detection results as well as threshold-independent metrics such as ROC-AUC and PR-AUC. Table 3 summarizes the results.

The detector comparison revealed heterogeneous behavior across attack families. For FGSM, ANFIS achieved the best overall performance, reaching 76.92% accuracy, 0.786 ROC-AUC, and 0.769 F1-score. A similar trend was observed for

PGD, where ANFIS again obtained the strongest results, with 75.00% accuracy, 0.861 ROC-AUC, and 0.877 PR-AUC. These results indicate that the neuro-fuzzy detector can be competitive under moderate attacks, especially when the perturbation induces gradual rather than sharply separable changes in the classifier outputs.

For DeepFool, however, the best performance was obtained by RF, which achieved perfect separation in the current sample. ANFIS remained competitive in AUC-based evaluation, with 0.857 ROC-AUC and 0.882 PR-AUC, but did not surpass RF in accuracy. This result suggests that adversarial EEG detection is highly attack-dependent and that no single detector consistently dominates across all perturbation regimes. In practical terms, this supports a more balanced interpretation of ANFIS: rather than universally outperforming classical methods, it should be viewed as an uncertainty-aware complementary detector that may be particularly useful when continuous risk scoring is preferred over hard binary decisions.

Table 3. Performance of adversarial detectors across attack families. Best result per attack is shown in bold.

Attack	Detector	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
FGSM	RF	0.692	0.750	0.500	0.600	0.667	0.686
	SVM	0.385	0.375	0.500	0.429	0.405	0.541
	KNN	0.538	0.500	0.500	0.500	0.560	0.632
	ANFIS	**0.769**	**0.714**	**0.833**	**0.769**	**0.786**	0.675
PGD	RF	0.500	0.500	0.500	0.500	0.500	0.549
	SVM	0.417	0.429	0.500	0.462	0.389	0.479
	KNN	0.500	0.500	0.500	0.500	0.458	0.500
	ANFIS	**0.750**	**0.714**	**0.833**	**0.769**	**0.861**	**0.877**
DeepFool	RF	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	SVM	0.615	0.556	0.833	0.667	0.905	0.933
	KNN	0.615	0.556	0.833	0.667	0.667	0.562
	ANFIS	0.692	0.667	0.667	0.667	0.857	0.882

4.3 Uncertainty-Aware Analysis and Feature Relevance

The precision-recall behavior of ANFIS under DeepFool supports the usefulness of a continuous adversarial risk score. Even when it did not outperform RF in accuracy, ANFIS achieved 0.882 PR-AUC, indicating that its continuous output still captures relevant adversarial structure. This is particularly relevant in safety-critical BCI scenarios, where a graded suspicion score may be more informative than a rigid binary prediction. As shown in Fig. 2, the precision–recall curve remains in a high-performance region for a broad range of thresholds,

reinforcing the ranking ability of the ANFIS score under DeepFool. In addition, Fig. 3 shows that clean and adversarial samples tend to occupy different confidence regions, although overlap is still visible. This behavior is consistent with the idea that adversarial perturbations in EEG do not always create a perfectly separable signature, but rather induce gradual shifts in uncertainty-related quantities derived from the classifier output.

To better understand the contribution of the extracted features, we also performed a simple ablation study with Random Forest on the FGSM scenario. The results in Table 4 show that the pair formed by maximum confidence and logit margin achieved the best performance, with 76.92% accuracy, 0.798 ROC-AUC, and 0.863 PR-AUC. This finding suggests that much of the detector's discriminative capacity comes from uncertainty-related quantities derived from the classifier output, particularly those associated with decision sharpness and boundary proximity. By contrast, adding all five features did not improve performance, indicating that feature relevance is uneven and that compact uncertainty-oriented subsets may be more effective than larger feature vectors.

Table 4. Feature ablation with Random Forest for FGSM detection.

Feature Set	Accuracy	ROC-AUC	PR-AUC
All features	0.692	0.667	0.686
Entropy + Confidence + Margin + Stability	0.692	0.643	0.720
Norm + Confidence + Margin + Stability	0.538	0.738	0.749
Norm + Entropy + Margin + Stability	0.692	0.667	0.686
Norm + Entropy + Confidence + Stability	0.615	0.548	0.526
Norm + Entropy + Confidence + Margin	0.692	0.690	0.710
Entropy + Confidence + Margin	0.692	0.762	0.843
Confidence + Margin	**0.769**	**0.798**	**0.863**

4.4 Limitations

Despite these promising findings, it is important to interpret the results with some caution, as they are still preliminary. One key limitation is the relatively small number of effective adversarial examples retained after filtering: only 21 for FGSM, 19 for PGD, and 21 for DeepFool. With such limited sample sizes, the evaluation becomes more sensitive to variation, which can influence performance metrics. This likely contributes to the perfect separation achieved by the Random Forest in the DeepFool scenario, a result that may not hold under larger or more diverse datasets.

In addition, the study was conducted under a very specific setting, involving a single subject, one EEG dataset, and one backbone architecture. While this is common in early-stage investigations aimed at exploring feasibility, it

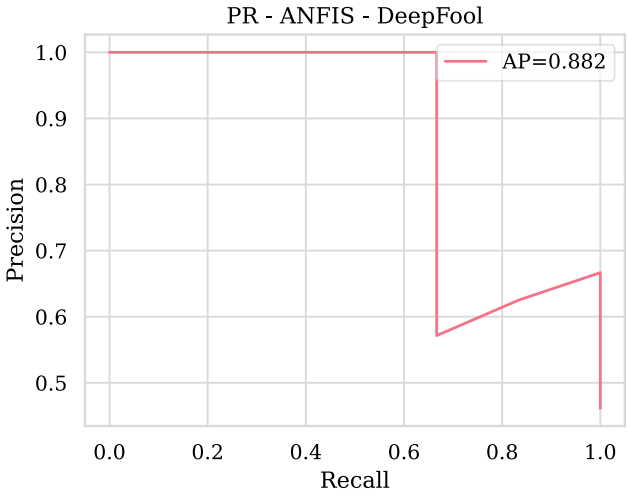

Fig. 2. Precision–recall curve of ANFIS under DeepFool, with AP = 0.882.

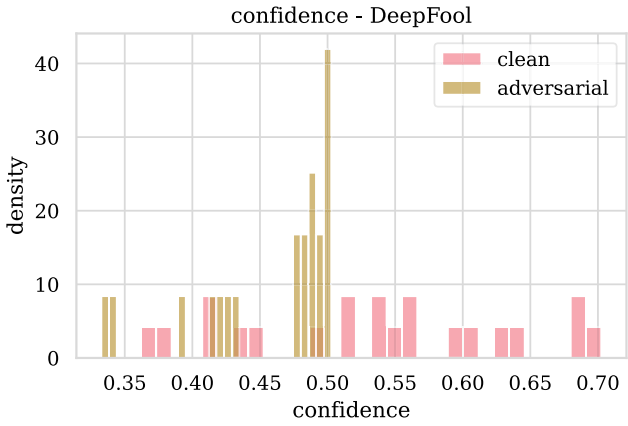

Fig. 3. Distribution of maximum confidence for clean and adversarial DeepFool samples.

restricts the ability to generalize the findings. In practical terms, it remains unclear whether the same performance would be observed across different individuals, experimental paradigms, or model architectures.

For this reason, further validation is necessary. Future work should expand the analysis to include multiple subjects, a wider range of EEG datasets, and different modeling approaches. Only with this broader evaluation will it be possible to determine whether the patterns observed here are robust and generalizable to more realistic and varied scenarios.

5 Conclusion

This paper investigated adversarial robustness in EEG-based Brain–Computer Interfaces by combining attack emulation and adversarial detection analysis. Using EEGNet as a baseline classifier, we showed that FGSM, PGD, and DeepFool consistently reduce classification accuracy, confirming the vulnerability of EEG-based pipelines to adversarial perturbations.

We then compared classical machine learning detectors with a neuro-fuzzy alternative based on ANFIS. The results showed that detector performance is strongly attack-dependent: ANFIS obtained the best results for FGSM and PGD, while Random Forest achieved the strongest performance under DeepFool in the current experimental setup. These findings reinforce the main conclusion of this study: adversarial detection in EEG-based BCIs is attack-dependent, and no single detector should be assumed to generalize across perturbation regimes without broader validation.

Beyond hard classification results, ANFIS remained competitive in AUC-based evaluation and offered an interpretable continuous adversarial risk score. This characteristic is especially relevant in safety-critical BCI scenarios, where representing uncertainty may be more useful than strict binary decisions. In addition, the feature ablation analysis indicated that confidence and logit margin are particularly informative for adversarial detection, highlighting the relevance of uncertainty-related quantities derived from classifier outputs.

The present study remains preliminary due to the limited number of effective adversarial examples and the restriction to a single subject, one EEG dataset, and one backbone architecture. Future work should therefore extend the analysis to cross-subject settings, additional EEG paradigms, and richer representation strategies, as well as further explore interpretable neuro-fuzzy mechanisms for adversarial risk estimation in BCI systems.

Acknowledgements. The authors would like to thank FAPERGS (24/2551-0001396-2, 25/2551-0002584-2), CNPq (307416/2025-9), FAPERGS/CNPq (23/2551-0000126-8) and the Fesurv-University of Rio Verde.

References

1. Aissa, N.E.H.S.B., Lakas, A., Korichi, A., Kerrache, C.A., Belkacem, A.N.: 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), pp. 156–161 (2023). https://doi.org/10.1109/IIT59782.2023.10366492
2. Alsereidi, M., Awadallah, A., Alkaabi, A., Yoon, S., Yeun, C.Y.: Data poisoning against federated learning: Comparative analysis under label-flipping attacks and GAN-generated EEG data. In: 2024 2nd International Conference on Cyber Resilience (ICCR), pp. 1–5 (2024). https://doi.org/10.1109/ICCR61006.2024.10533012

3. Alshebli, S., Alshehhi, M., Yeun, C.Y.: Investigating how data poising attacks can impact an EEG-based federated learning model. In: 2024 2nd International Conference on Cyber Resilience (ICCR), pp. 1–6 (2024). https://doi.org/10.1109/ICCR61006.2024.10532875
4. Bernal, S.L., Celdrán, A.H., Pérez, G.M., Barros, M.T., Balasubramaniam, S.: Security in brain-computer interfaces: state-of-the-art, opportunities, and future challenges. ACM Comput. Surv. **54**(1) (2021). https://doi.org/10.1145/3427376
5. Bidgoli, N.S., Kheradpisheh, S.R., Farahani, H.: EEG anomaly detection using generative adversarial networks (GANs). In: 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), pp. 1–6 (2024). https://doi.org/10.1109/AISP61396.2024.10475243
6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
7. Brunner, C., Leeb, R., Müller-Putz, G.R., Schlögl, A., Pfurtscheller, G.: BCI competition 2008 graz data set a (2008). https://www.bbci.de/competition/iv/desc_2a.pdf
8. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
9. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (1967). https://doi.org/10.1109/TIT.1967.1053964
10. Dalmazo, B.L., Vilela, J.P., Curado, M.: Performance analysis of network traffic predictors in the cloud. J. Netw. Syst. Manage. **25**, 290–320 (2017)
11. Jang, J.S.: ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybern. **23**(3), 665–685 (1993). https://doi.org/10.1109/21.256541
12. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for EEG-based brain–computer interfaces. J. Neural Eng. **15**(5), 056013 (2018). https://doi.org/10.1088/1741-2552/aace8c
13. Leite, L., Santo, Y., Dalmazo, B., Riker, A.: 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, pp. 794–800. SBC, Porto Alegre, RS, Brasil (2024). https://doi.org/10.5753/sbseg.2024.241680, https://sol.sbc.org.br/index.php/sbseg/article/view/30070
14. Li, Y., Yu, X., Yu, S., Chen, B.: 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), pp. 1–6 (2022). https://doi.org/10.1109/MLSP55214.2022.9943479
15. Naik, N., Jenkins, P., Savage, N., Yang, L., Boongoen, T., Iam-On, N.: Fuzzy-import hashing: a malware analysis approach. In: 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8 (2020). https://doi.org/10.1109/FUZZ48607.2020.9177636
16. Pandey, S.K., Mustafa, K., Ahson, S.I.: A checklist based approach for the mitigation of buffer overflow attacks. In: 2007 Third International Conference on Wireless Communication and Sensor Networks, pp. 115–117 (2007). https://doi.org/10.1109/WCSN.2007.4475759
17. Quevedo, A., Ayres, D., Dimuro, G., Riker, A., Lucca, G., Dalmazo, B.L.: Optimizing big data traffic prediction using generalizations of choquet integral with adaptive weighting. In: ICC 2025 - IEEE International Conference on Communications, pp. 4872–4877 (2025). https://doi.org/10.1109/ICC52391.2025.11161751
18. Rauber, J., Zimmermann, R., Bethge, M., Brendel, W.: Foolbox native: fast adversarial attacks to benchmark the robustness of machine learning models in pytorch,

tensorflow, and jax. J. Open Source Softw. **5**(53), 2607 (2020). https://doi.org/10.21105/joss.02607
19. Sanei, S., Chambers, J.A.: Machine learning for EEG analysis, pp. 177–233 (2021). https://doi.org/10.1002/9781119386957.ch7
20. Schalk, G., McFarland, D., Hinterberger, T., Birbaumer, N., Wolpaw, J.: BCI 2000: a general-purpose brain-computer interface (BCI) system. IEEE Trans. Biomed. Eng. **51**(6), 1034–1043 (2004). https://doi.org/10.1109/TBME.2004.827072
21. Shi, Y., Sagduyu, Y., Grushin, A.: How to steal a machine learning classifier with deep learning. In: 2017 IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–5 (2017). https://doi.org/10.1109/THS.2017.7943475
22. Tan, W., et al.: Noisecam: explainable AI for the boundary between noise and adversarial attacks. In: 2023 IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1–8 (2023). https://doi.org/10.1109/FUZZ52849.2023.10309766
23. Zhai, J., Song, G., Zhou, T.: Recognition of epilepsy EEG by two-view adversarial-incentived-based t-s fuzzy classifier. In: 2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR), pp. 146–153 (2023). https://doi.org/10.1109/AIHCIR61661.2023.00032
24. Zhang, H., Li, L., Qiao, L.: A control and attack detection scheme for fuzzy systems against cyber-attacks. In: 2023 IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1–6 (2023). https://doi.org/10.1109/FUZZ52849.2023.10309687