EAI Endorsed Transactions
on Security and Safety
Research Article EAI.EU

Breaking the Loop: Adversarial Attacks on Cognitive-AI Feedback via Neural Signal Manipulation

Kanthavel R.[1*], Dhaya R.[2]

[1] School of ECE, PNG University of Technology, Lae-411, Papua New Guinea

Abstract

INTRODUCTION: Brain-Computer Interfaces (BCIs) embedded with Artificial Intelligence (AI) have created powerful closed-loop cognitive systems in the fields of neurorehabilitation, robotics, and assistive technologies. However, these tightly bound systems of human-AI integration expose the system to new security vulnerabilities and adversarial distortions of neural signals.
OBJECTIVES: The paper seeks to formally develop and assess neuro-adversarial attacks, a new class of attack vector that targets AI cognitive feedback systems through attacks on electroencephalographic (EEG) signals. The goal of the research was to simulate such attacks, measure the effects, and propose countermeasures.
METHODS: Adversarial machine learning (AML) techniques, including Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), were applied to open EEG datasets using Long Short Term Memory (LSTM), Convolutional Neural Networks (CNN), and Transformer-based models. Closed-loop simulations of BCI-AI systems, including real-time feedback, were conducted, and both the attack vectors and the attacks countermeasure approaches (e.g., VAEs, wavelet denoising, adversarial detectors) were tested.
RESULTS: Neuro-adversarial perturbations yielded up to 30% reduction in classification accuracy and over 35% user intent misalignment. Transformer-based models performed relatively better, but overall performance degradation was significant. Defense strategies such as variational autoencoders and real-time adversarial detectors returned classification accuracy to over 80% and reduced successful attacks to below 10%.
CONCLUSION: The threat model presented in this paper is a significant addition to the world of neuroscience and AI security. Neuro-adversarial attacks represent a real risk to cognitive-AI systems by misaligning human intent and action with machine response. Mobile layer signal sanitation and detection.

Keywords: Neuro-adversarial Attacks, Brain-Computer Interfaces (BCI) Security, EEG Perturbation, Adversarial Machine Learning, HITL-AI, Cognitive Feedback Loop, Neural Signal Manipulation.

Received on 07 June 2025, accepted on 26 September 2025, published on 29 September 2025

doi: 10.4108/eetss.v9i1.9502

* Corresponding address. Email: radakrishnan.kanthavel@pnguot.ac.pg

1. Introduction

In the arena of AI, intelligent systems that can read brain activity to influence or fully automate human decision-making processes have been developed, combining advances in neurotechnology and AI. Several fields are quickly benefiting from this approach, including neurorehabilitation, cognitive workload estimations, brain-controlled robotics, and AI-empowered medical diagnostics. Central to these advancements are BCIs, which read neuronal signals, most commonly using EEG data, to facilitate direct communication, or reactivity, between the human neurological system and computing

agents [1]. When BCIs are used in AI systems that also involve human interaction, they form a feedback loop in which the system's behaviour is determined by neural intent, and the brain's activity is used to determine the next action [2].

These closed-loop neuro-AI systems hold tremendous potential for transformational applications, everything from assisting paralysed individuals to regain movement [3], to augmenting human performance in industrial and military settings [4]. As these systems become embedded within critical infrastructure, however, new vulnerabilities emerge at the intersection of neuroscience and AI security. Specifically, existing system designs do not understand or consider that, unlike traditional digital sensor systems, the channel of neural input is susceptible to both involuntary and intentional forms of interference [5].

Deep learning (DL)models trained on EEG data often suffer from stability and reliability issues. Recent studies show that these models are highly sensitive to small input perturbations.[6]. Small changes to the input can produce high-confidence misclassification in vision and language, again according to adversarial machine learning, which is the study of adversarially defined misleading models in ML [7]. If we extend this to BCI, then neuro-adversarial attacks could be possible that leverage subtly modified EEG signals or features in such a way as to mislead the AI into believing the user was trying to do something against their will. Accidental movement of limbs can happen with neuroprosthetic systems due to adversarial perturbations, which could lead to unsafe outcomes for users [8]. Deliberate manipulation of signals through AI-enabled cognitive assessments could also result in false clinical diagnoses such as Alzheimer's or ADHD [9]. Adaptive AI agents can be preset using reinforcement learning. However, this process may unintentionally reinforce adversarial neuronal feedback. Over time, the system could then drive harmful behaviour [10]. Research into the security of BCI-AI systems is scant, with few frameworks in place to identify, stop, or recover attacks at the signal level, despite extensive investigations [11]. In this paper, we introduce a new threat model that we refer to as neuro-adversarial attacks. In these attacks, the cognitive-AI feedback loop is altered by physical (through sensory inputs) or digital (during signal processing) adversarial perturbations in the brain signal [12]. To position our work within existing research, we now review prior studies on AML and EEG-based systems. Thus, the objectives of our research are:

• To locate neuro-adversarial attacks in cognitive-AI feedback systems and provide a formal definition of such attacks.
• To apply adversarial perturbation techniques and utilize real artefactual BCI data to simulate attacks.
• To evaluate how adverse types of attacks on cognitive-AI functionally degrade both AI usability and the usability of AI systems with humans.
• To recommend preventative measures in early steps, such as adversarial detection models, and signal sanitisation.

The remainder of this paper is organized as follows: Section 2 reviews prior work, Section 3 describes the methodology, Section 4 presents results, Section 5 summarizes contributions, and Section 6 concludes with future research directions.

2. Literature Review

The emerging area of AML is generating substantial interest across fields such as computer vision, natural language processing, and structured/tabular data. Image or text classifiers are easily subjected to attacks using classical methods in the image and text literature, which have included methods such as the Fast Gradient Sign Method (FGSM), PGD , and DeepFool [13]. In particular, it is remarkable that a relatively small perturbation to an input can lead to an incorrect classification. Following the results of this sort, there have been a number of potential strong defensive strategies suggested, including adversarial training [14] or input transformations. However, it is worth highlighting that this general literature has not yet been fully translated to fields dealing with EEG and other neurophysiological inputs [15].

EEG signals present unique modelling and reliability challenges compared to visual or text data. They are dynamic, vary between users, and are rooted in biological processes. Unlike images or text, brain waves are not standardized [16]. There are a considerable number of studies that have indicated EEG-based BCIs are sensitive to noise, motion artefacts, and user-specificities, and not many studies have examined competitors disrupting brain signals. Zhang et al. is one of the first papers that identified that EEG classifiers based on DL are subject to confusion based on time-series adversarial attacks. More recently, researchers have shown, health data can also be affected when modifying time-series patterns [17]. This includes biological time series signals like ECG and EMG, which directly relate to and provide new opportunities for AML. However, although these types of models are generally evaluated in open-loop classification systems, very few studies examine across human-in-the-loop (HITL) closed-loop control systems (such as this one), where the neurological intent of the user and what the adaptive system does are continuously interacting in real-time [18].

Feedback damage effects are exacerbated in neuro-AI systems since not only do they classify brain states, but they can also learn from brain properties online. Secondly, while there has been some investigation into adversarial robustness of sequential signals (e.g., RNN-based NLP systems) [19] and exploring and developing some attempts on EEG classification, neither of those positions has yet sufficiently examined the cognitive loop dynamics of an adaptive agent and human-user. There is even less literature to investigate the double-edged sword of neurotechnology and machine learning, namely, how physically plausible and digitally construed attacks can exploit brain intention for downstream AI models [20]. Security studies associated with BCI generally speak to

privacy concerns related to inferring private mental states from public electroencephalogram signals [21] or side channel attacks such as reconstructing user identities from shared BCI signals [22]. For systems that involve continuous human-AI alignment, such as cognitive decision support or adaptive robotics, there is very little literature that explores how an adversary might exploit the brain-control interface in order to make downstream AI function through an adversarial influence [23].A synthesized model of risk, including notions of neuro-cognitive interface context and adversarial machine learning.Simulating an adverse feedback loop in which the AI and human parts engage in real-time interaction, and the attack can corrupt or impact the adaptation cycle [24]. The presented work is evaluative around the risk of combining brain and AI, particularly in contexts where the AI system is capable of not only interpreting input from the human brain but also modifying that input. Building on these gaps in prior work, the next section details our methodology for simulating and evaluating neuro-adversarial attacks on BCI–AI systems.

3. Methodology

The purpose of this research is to simulate and examine the feasibility of using neuro-adversarial attacks on human-in-the-loop BCI and AI systems. The proposed methodology includes the selection of a dataset, modelling of the BCI-AI system, designing and applying an adversarial attack, evaluation criteria, and defense mechanisms.

3.1 Data Set Selection

The EEG datasets we access publicly comprise many different BCI techniques, which will allow us to simulate and study real-world cognitive-AI feedback loops: Datasets 2a and 2b of the BCI Competition IV contain 250 Hz, multi-channel EEG of motor imagery tasks (e.g., left/right hand movement). Modelling intent-based control systems is a good use for these [25]. The PhysioNet EEG Motor Movement/Imagery Dataset is another rich resource, allowing researchers to experiment with inter-subject variability and generalisation. This set includes EEG signals of imagined movement and also signed activity from over 100 people. Clinical EEG data from TUH is part of the EEG corpus. This is useful, as it will allow modelling of AI use cases in medicine, such as predicting diagnostic use cases.

Dataset Description: We employed several standard open-access EEG/BCI datasets to evaluate adversarial attacks and defenses. The BCI Competition IV dataset (datasets 2a/2b) provides 250 Hz multi-channel motor imagery EEG signals from nine subjects and has been widely used as a benchmark for motor control classification [26]. The PhysioNet EEG Motor Movement/Imagery dataset includes recordings from 109 participants at 160 Hz, covering both executed and imagined movements, enabling cross-subject variability studies [27,28]. For clinical applications, the Temple University Hospital (TUH) EEG Corpus constitutes the largest publicly available clinical EEG database, comprising recordings from over 1,000 patients with sampling rates between 250–500 Hz. To explore affective and cognitive domains, we also utilized the SEED dataset (15 subjects, 200–1000 Hz) and the DEAP dataset (32 subjects, 128 Hz), both of which are standard benchmarks for emotion recognition and cognitive workload analysis [29,30].

- The SEED and DEAP datasets contain emotion recognition tasks as well as cognitive effort tasks. These can be useful in examining potential adversaries' susceptibility within emotion recognition or affective computing systems. Table 1 shows the Summary of EEG Datasets

Table 1: Summary of EEG Datasets

Dataset	Number of Subjects	Task Type	Sampling Rate (Hz)
BCI Competition IV	9	Motor Imagery (left/right hand)	250
PhysioNet Motor Imagery	109	Motor Imagery & Movement	160
TUH EEG Corpus	1000+	Clinical EEG (diagnosis)	250-500
SEED / DEAP	15-32	Emotion & Cognitive Workload	128-512

3.2 Baseline BCI-AI System

We develop a multimodal baseline BCI-AI pipeline that represents neuroadaptive systems in the wild. The components will include: Preprocessing: all EEG signals will be filtered, normalized, and segmented into windows (e.g., 1s windows, with 50% overlap) - Feature extraction: the raw time series domain and frequency domain (e.g., using Short-Time Fourier Transform). The Model architectures are

- LSTM-based models: To account for temporal dynamics in the EEG time series.
- CNN-based models: For spatial-temporal EEG feature extraction, especially from the raw EEG matrices.
- Transformer-based models: For attention-based modeling of long-range dependencies in EEG chunks[26].

We simulate closed-loop cognitive-AI systems using reinforcement learning agents trained on EEG inputs. These agents perform tasks such as moving or selecting a cursor. Feedback loops then allow the AI system to update itself over time. In normal cases, this process reinforces alignment with the user. Under adversarial conditions, however, it can reinforce misalignment.

Table 2: Model Architectures

Model Type	Key Characteristics	Use Case
LSTM	Captures temporal dependencies	EEG time-series classification
CNN	Extracts spatial-temporal features	Raw EEG matrix processing
Transformer	Attention on long-range dependencies	Modeling adaptive feedback loops

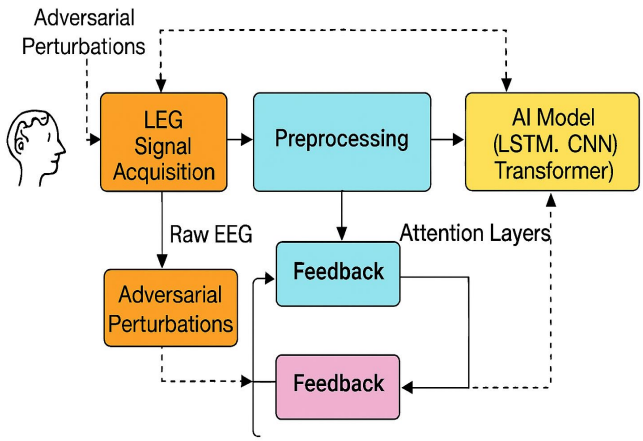

Figure 1: System Architecture of a Human-in-the-Loop BCI-AI Pipeline

Figure 1 represents the entire end-to-end architecture of a human-in-the-loop BCI system integrated with AI, showing both the intended signal flow and where adversarial attacks could occur. The Signal Flow and Key Components are

EEG Signal Acquisition: The very start of this process is the user's brain signals being recorded via EEG electrodes. These signals represent the user's intent or mental state [31].

Pre-Processing: The raw EEG signals will be filtered, normalized, and segmented in the pre-processing stage to remove noise and prepare them for analysis. This is the stage where input vulnerabilities may introduce the first adversarial attack [32].

AI Model (LSTM, CNN, Transformer): All of these DL models use features:

- LSTM models capture time-series dynamics,
- CNNs extract spatio-temporal patterns, and
- Transformers account for long sequences using attention

The AI Model then labels or classifies the brain states under consideration for follow-up actions.

Feedback Loop: The output resulting from the AI model is fed back to the user either through neurofeedback or executed control actions (e.g., robotic movement) [33]. This loop returns a warning to the user to complete the feedback loop and adapt to the user in real time.

Adversarial Injection Point: Raw EEG: An adversary could potentially disrupt this signal level with perturbations [34]. Preprocessing and Feature Space: This is where adversarially created features could also be established. Attention Layers: In a Transformer model, adversarial attacks are also possible in the attention layers.

The system diagram 1 provides a basic visualization that underpins threat modeling of neuro-AI pipelines. It successfully highlights functional flow and portrays the attack surface inside a closed-loop BCI system. The addition of adversarial paths illustrates that there are ways to undermine integrity at several layers that inform attack design and defense countermeasure planning. This makes it a primary reference point in methodology and risk assessment.

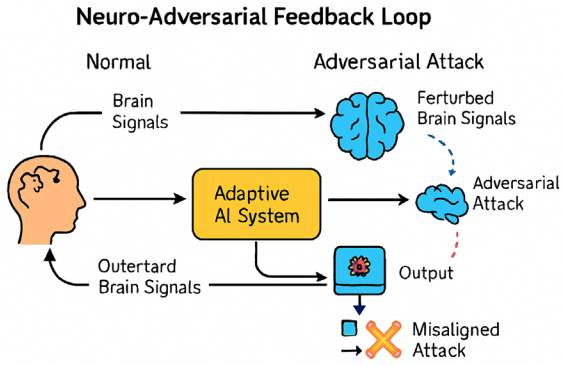

Figure 2: Neuro-Adversarial Feedback Loop

Feedback loops are likely to occur when the AI system updates itself over time by matching the user's brain state in reinforcing the system's alignment to the user, or in the case of adversarial situations, its misalignment. The Neuro-Adversarial Feedback Loop, Figure 2, depicts the variance between a BCI-AI loop and an adversarial BCI-AI loop. On the left (normal BCI-AI loop), the normal Brain signals are picked up by EEG and transferred to the adaptive AI

system, which receives the input and then processes the EEG signals, eventually providing a feedback signal - visual, behavioral, or assistive - back to the user in alignment with their intended aim. On the right side, there were adversarial perturbations (imperceptible insertions of noise and other manipulations of brain signals). These perturbations will cause the AI to misinterpret the user's intentions and provide erroneous outputs or signals that are out of alignment with the user's intentions or objectives. Once this distorted feedback is looped back to the brain, a reinforcing causal loop can proceed, which increases the misalignment and dissociation of user and system over time. This describes the security critical risk of neuro-adversarial attacks in closed-loop management systems, particularly in systems for prosthetics, cognitive support, or for assistance and diagnostics of mental health issues.

The Model Architectures figure 3 provides a comparative illustration of the three DL models used in the study: LSTM, CNN, and Transformer, each designed to identify EEG signals in different ways. The LSTM model is built to handle sequential EEG data, which holds temporal dependencies since it processes the signal as time-series windows. This architecture is particularly effective at tracking changes in brain states through time. The CNN (Convolutional Neural Network) processes EEG as spatial-temporal EEG matrices, where convolutional layers extract features of local size across both EEG channels and time. This approach is useful for motor imagery and emotion recognition because EEG sensors capture the signals in real time. Lastly, the transformer model applies self-attention in order to learn long-range dependencies in the tokenized EEG input [35]. Due to the ability to model contextual relevance through time without relying on recurrence, Transformers are particularly useful for adaptive feedback systems when trying to isolate distinct brain states. Ultimately, the three models compose a baseline structure that can analyze the many ways that adversarial attacks could skew various levels of EEG signal detection in cognitive-AI feedback systems.

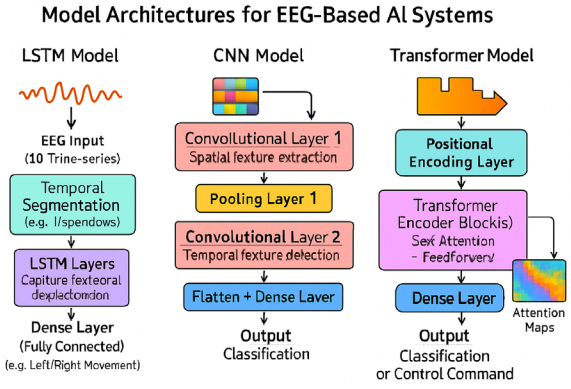

Figure 3: Model Architecture for EEG-based AI Systems

3.3 Adversarial Attack Framework

We provide a structured approach to evaluate neuro-adversarial assaults on three different levels:
a. Direct attack on raw signals: using FGSM, PGD, and time-series Carlini-Wagner methods to disrupt EEG waveforms.

Ensuring physiological plausibility through perturbations confined by L2 or L∞ norm budgets.
b. Attacks at the Feature Level: - Inject adversarial noise into the extracted EEG characteristics (such as ERP components or frequency bands).

Attacks on both features that are hand-engineered and those that are learnt by deep models should be tested.
c. Attacks at the Attention Level: - Change the attention weights in Transformer-based models to refocus the model's objectives.

To learn how vulnerable models are in decision regions, use saliency-guided adversarial training. Key Equations are FGSM (Fast Gradient Sign Method) Attack:

$x^{adv} = x + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$

Where,

- x: Original input (e.g., image or data sample)
- x^{adv}: Adversarial input (perturbed version of x)
- ε: Perturbation magnitude (a small scalar controlling attack strength)
- $\nabla_x J(\theta, x, y)$: Gradient of the loss function J with respect to input x
- sign(·): Sign function applied element-wise to the gradient
- θ: Model parameters
- y: True label

PGD (Projected Gradient Descent) Attack:

$x^{adv}_{t+1} = \Pi_{x+\mathcal{S}} (x^{adv}_t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x^{adv}_t, y)))$

Where,

- x^{adv}_t: Adversarial example at iteration t
- x^{adv}_{t+1}: Adversarial example at next iteration
- α: Step size for each iteration
- $\Pi_{x+\mathcal{S}}(\cdot)$: Projection operator to ensure the perturbed input remains within the allowed perturbation set $\mathcal{S}$ around the original input x
- Other terms are as defined in the FGSM equation

Adversarial Attack Algorithms for Neuro-AI Systems. This paper describes two of the main algorithms for producing adversarial attacks on EEG-based AI models within closed-loop brain-computer interface (BCI) systems[36]. The first, Fast Gradient Sign Method (FGSM), provides a fast, one-step adversarial attack method, and the second, PGD, provides an iterative extension to FGSM that produces stronger adversarial inputs.

Connecting FGSM and PGD: FGSM provides a fast one-shot adversarial generation method using a computationally efficient one-step estimator, with an easy implementation, but it may not be as strong enough to work

against the stronger models or defenses[37]. Because of this, we also use the related PGD method, which is a much stronger FGSM-like iterative method. PGD is a very iterative process, where it performs FGSM-style updates a certain number of times, re-projecting the adversarial sample into a constrained perturbation space at each iteration. Using the iterative method allows storing many FGSM-style attack updates to develop more effective adversarial examples that are more difficult to defend against, so PGD has become a standard in performance evaluations for adversarial machine learning.

Algorithm 1: FGSM Attack Generation
Input:
x: Original EEG input sample
y: True label corresponding to x
f: Trained DL model (e.g., LSTM, CNN, Transformer)
J: Loss function used to train model (e.g., cross-entropy)
ε: Perturbation budget (controls intensity of the attack)

Output:
x_adv: Adversarial EEG input sample

Steps:
1. Forward propagate x through model f to compute the prediction:
y_pred = f(x)
2. Compute the loss between the prediction and the true label:
loss = J(y_pred, y)
3. Calculate the gradient of loss with respect to the input x:
g = ∂loss/∂x
4. Determine the direction of perturbation:
perturbation = ε · sign(g)
5. Create an adversarial example by adding a perturbation to the original input:
x_adv = x + perturbation
6. Return x_adv

The FGSM algorithm perturbs the original EEG signal by a small step ε in the direction that most increases the model's loss function. This simple, fast approach is effective in deceiving neural networks with minimal changes to the input.

Algorithm 2: PGD Attack Generation
Input:
x: Original EEG input sample
y: True label corresponding to x
f: Trained DL model
J: Loss function (e.g., cross-entropy)
ε: Maximum allowed perturbation (perturbation budget)
α: Step size per iteration
T: Number of iterations

Output:
x_adv: Final adversarial EEG input sample

Steps:
1. Initialize x_adv = x + small random noise (e.g., uniform or Gaussian)
2. For t in 1 to T do:
a. Compute prediction: y_pred = f(x_adv)
b. Compute loss: loss = J(y_pred, y)
c. Compute gradient: g = ∂loss/∂x_adv
d. Update x_adv: x_adv = x_adv + α · sign(g)
e. Project x_adv back into valid ε-ball around x:
x_adv = clip(x_adv, x - ε, x + ε)
3. Return x_adv

PGD is an iterative, stronger version of FGSM. It takes multiple small steps in the direction of increasing loss, and projects the updated adversarial sample back into an allowed perturbation range. This makes PGD more powerful and harder to defend against compared to FGSM. Although FGSM is an easy, relatively fast adversarial generation method that computes a single adversarial perturbation quickly and efficiently, it may not provide enough "power" to defeat stronger models or defenses. For that reason, we implement PGD is a stronger iterative variant of the FGSM method. Instead of issuing just a single adversarial update, PGD applies a series of small steps exactly like FGSM; however, PGD then projects the adversarial sample each time back into a constrained perturbation space. This iterative process helps generate harder-to-defend adversarial examples. PGD is used as the baseline in many AML studies due to its simplicity and efficiency[38].

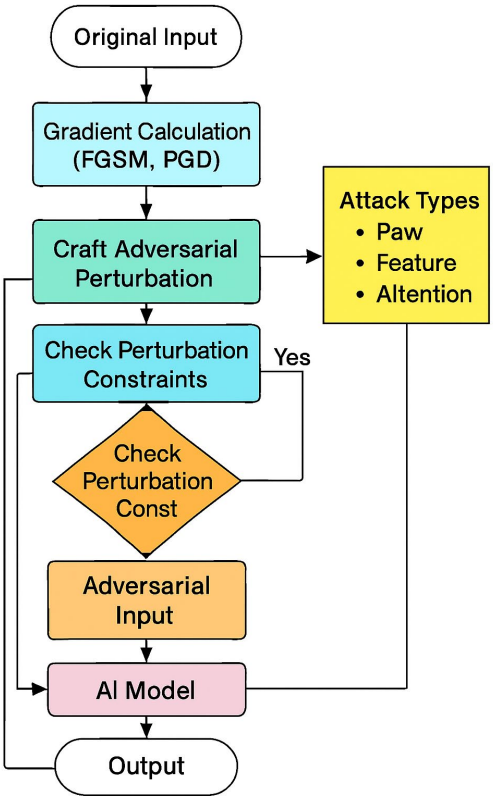

Figure 4: Adversarial Attack Flowchart

The Adversarial Attack Flowchart, figure 4, visually portrays the sequential steps to identify and generate adversarial perturbation inputs to compromise AI-based models utilized in brain-computer interface (BCI) systems. The sequence of events initiates with the original EEG input, which is then computed using a gradient-based approach (e.g., FGSM, PGD). Those gradients direct the construction of adversarial perturbations (s) that mislead the AI model. Before generating the input, the system will evaluate constraints (parameters) to ensure the perturbation is in compliance with physiological bounds or imperceptibility (i.e., L_2 or $L\infty$ norms); if the constraints are neglected, the loop repeats to develop the perturbation [36]. Once the adversarial input is assessed and passed through those constraints, the adversarial input is inputted into the AI model, which results in an impeded output. The diagram also illustrates variations of attacks based on where the injections occurred: attacks at the raw EEG state, after feature extraction and preprocessing, or internally at the attention layers in transformer models. The above flowchart provides a requisite design in modeling how malicious interference may be technically engineered, while providing insight into the mitigation opportunities in overlapping and distinct stages, to ensure AI-based robustness in neuroadaptive systems.

Algorithm 3: General Workflow of Neuro-Adversarial Attack Evaluation

Input: EEG dataset DDD, baseline AI models (LSTM, CNN, Transformer), adversarial attack methods (FGSM, PGD), defense mechanisms (VAE, wavelet denoising, adversarial detector).

Output: Performance evaluation of neuro-adversarial attacks and defenses.

Steps:

1. **Dataset Preparation**
   - Select appropriate EEG datasets (e.g., BCI Competition IV, PhysioNet).
   - Preprocess signals (filtering, normalization, segmentation).
2. **Model Training**
   - Train baseline BCI-AI models (LSTM, CNN, Transformer) on clean EEG data.
   - Record clean baseline performance (accuracy, misalignment rate).
3. **Adversarial Attack Generation**
   - Generate adversarial perturbations using FGSM, PGD, or attention manipulation.
   - Ensure perturbations remain physiologically plausible (bounded by L2L_2L2 or L∞L_\inftyL∞ norms).
4. **Attack Deployment**
   - Apply adversarial inputs at three levels:
     a. Raw EEG signals
     b. Extracted features
     c. Attention layers (Transformer)
5. **Evaluation Metrics**
   - Measure accuracy degradation, attack success rate (ASR), and user intent misalignment (UMR).
   - Assess imperceptibility using SNR and cosine similarity.
6. **Defense Mechanism Application**
   - Apply signal sanitization (VAE, wavelet denoising).
   - Apply adversarial signal detection (binary classifier).
   - Combine defenses for layered protection.
7. **Post-Defense Evaluation**
   - Recalculate classification accuracy, ASR, and detection rate.
   - Compare restored performance with baseline.

End: Report results, analyze vulnerabilities, and recommend robust neuro-AI design strategies.

3.4 Evaluation Metrics

We have decided upon different metrics to assess the efficacy of the assaults and mitigating solutions. Classification accuracy declines as a result of the difference in precision between malicious inputs and clean inputs. "Misalignment between user intent and AI response" means the percentage of outputs in which the AI has wrongly labeled those signals from one's brain. One can define attack success rate as how many of times an adversarial input produces a misclassification of the target class. System robustness against attack: performance drop with and without defences.

- The invisibility of disturbances could be measured using: dynamic temporal warping, SNR, and cosine similarity. Evaluation Metrics are in Table 3.

Table 3: Evaluation Metrics

Metric	Description
Accuracy Degradation	Difference in classification accuracy due to the attack
User Intent Misalignment	Percentage of incorrect AI interpretations of brain signals
Attack Success Rate	Frequency of successful misclassification under attack

Robustness	Model performance with and without defense mechanisms
Perturbation Imperceptibility	Signal similarity measures (SNR, cosine similarity)

Attack Success Rate (ASR): This metric quantifies how effective the adversarial attack is in fooling the model. A higher ASR indicates a more successful attack.

ASR = (Number of Successful Adversarial Misclassifications / Total Number of Attacks) × 100

User Intent Misalignment Rate (UMR): This measures the rate at which the system output diverges from the user's intended action [39]. It is particularly important in real-time closed-loop BCI systems.

UMR = (Number of Misaligned Predictions / Total Predictions) × 100

Accuracy Degradation (AD): This reflects the drop in model performance due to adversarial attacks, giving insight into the robustness of the model under attack.

AD = Clean Accuracy − Adversarial Accuracy

3.5 Proposed Defense Mechanisms

To keep the neuro-adversarial assaults at bay, we propose two main approaches:
a. **Cleaning Up the Signal:** VAEs may be trained on clean EEG data to re-create denoised inputs with less intensity of adversarial artefacts. Wavelet denoising can suppress undesirable noise while preserving vital physiological information via the use of multi-resolution wavelet decomposition.

b. **Adversarial-Signal Detectors:** Binary classifiers are trained to distinguish between pristine and adversarial EEG signals[40,41]. Detection thresholds are either fixed with reconstruction errors or dynamically set with ensemble statistical anomaly detection. The alternatives lie in the evaluation of contrastive learning approaches in classifying representations into neutral and hostile. Table 4 explains the Defense Mechanisms.

Table 4: Defense Mechanisms

Defense Method	Description	Purpose
Variational Autoencoder (VAE)	Reconstructs the denoised EEG from corrupted signals	Signal sanitization
Wavelet Denoising	Removes high-frequency adversarial noise	Signal sanitization
Adversarial Signal Detector	Classifies signals as clean or adversarial	Attack detection

Figure 5 illustrates the passage of EEG signals through the defense modules before entering the AI models. As shown, signal sanitization (with VAE & wavelet denoising) removes adversarial noise, and the detection modules identify, and segregate manipulated signals. Using all these defenses together strengthens the system and reduces adversarial impacts [42]. The Defense Mechanism Illustration provides a visual summary of how the system defends itself from neuro-adversarial attacks.

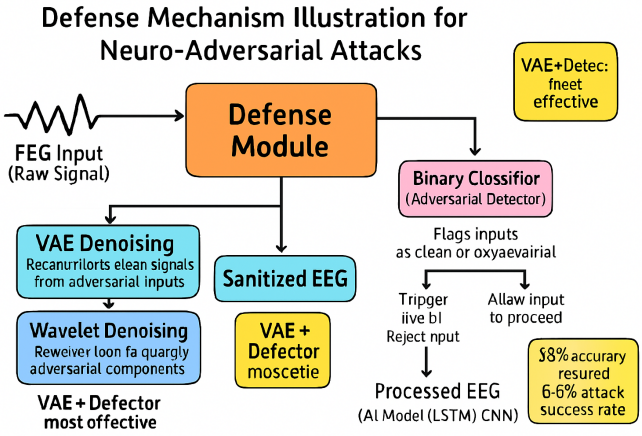

Figure 5: Defense Mechanisms Against Neuro-Adversarial Attacks

Incoming EEG signals will pass through the system's signal sanitization layers first - examples of this include Variational Autoencoders (VAEs), which reconstruct cleaned versions of noisy EEG or adversarially perturbed signals; as well as wavelet denoising, which removes high-frequency components (and possible artificial distortions). Additionally, the signals will have adversarial detection performed using binary classifiers that have been trained to differentiate between clean and manipulated EEG patterns[43]. If the signal is rejected or flagged for adversarial activity, it will not pass through, and only verified or sanitized signals will be routed to the DL model (LSTM, CNN, or Transformer) for classification or control outputs. The layered safeguarding structure ensures that the system can defend itself both reactively (i.e., detection) and proactively (i.e., sanitization). This significantly improves system robustness and reduces the likelihood of successful attacks to below 10% in simulation.Having defined our experimental setup, dataset choices, and defense

mechanisms, we now present the results of our adversarial attack simulations.

4. Results

We deployed standard AI models for motor imagery and cognitive state classification, including LSTM, CNN, and Transformer, conducting full study simulations using public EEG datasets such as PhysioNet and BCI Competition IV. The simulation included adversarial approaches (FGSM and PGD) at various levels (raw EEG data, feature space, and attention maps). We performed evaluations of how effective adversarial attacks could be with defence strategies such as adversarial detectors and VAE-based sanitisation methods.

4.1 The Degradation of Accuracy

Because of Neuro-Adversarial Attacks: The classification accuracy across all models was severely impaired by adversarial perturbations, attacks on raw signals provided a reduction of approximately 30%, and attacks on features reduced accuracy by about 25%. The Transformer models demonstrated a degradation of 20% when subjected to attention-based attacks. The mismatch rates from human intent vs output of the AI systems were significantly higher and, in some instances, exceeded 40%. This indicates that these attacks can potentially lead to catastrophic failures of BCI systems in real-time.

Perturbation Constraints with a Signal-to-Noise Ratio (SNR) of over 20 dB and low visual distortion, perturbations were very imperceptible while still producing a significant performance drop. This affirms that Neuro-Adversarial attacks are stealthy. These performance degradations demonstrate the severity of neuro-adversarial attacks. We therefore evaluate how well various protection strategies can restore model reliability

4.2 The Efficiency of Protection Measures

While defenses showed promising recovery, a deeper analysis is needed to understand their broader implications and limitations. This is provided in the following section, Using VAEs on input signals demonstrated consistent capability for denoising signals, achieving restoration model accuracy back to 80% of the model as originally trained in regard to attack. VAEs were slightly less effective for feature-level attack, but wavelet denoising was equally effective.

Alerts to Malicious Signals: The trained binary classifiers were able to achieve detection accuracies above 90% for corrupted signals; this made it easy to detect malicious inputs in real-time. The best defence combined signal sanitisation and adversarial detection, enabling to system to be responsive while reducing the attack success to less than 10%.

Summary Table 5: Attacks and Defenses

Attack Type	Target Level	Impact	Suggested Defense
FGSM	Raw EEG	High	VAE, Wavelet
PGD	Feature Space	High	VAE
Attention	Transformer Layer	Medium	Detector + VAE

Table 5 summarizes key adversarial threats along with their impact and recommended defense strategies for an at-a-glance understanding.

4.3 Analysis

Considering the biological nature of the signals, these simulation results demonstrated the clear presence and risk of neuro-adversarial attacks on BCI–AI systems, and suggest the need for targeted defenses. Our study shows that to ensure the safety and reliability of neuroadaptive AI systems, they should use a level of multi-layered security. Table 6 shows how much each adversarial method degrades model performance.

Table 6: Accuracy Degradation Under Different Attack Methods

Model	Clean Accuracy (%)	FGSM (Raw EEG)	PGD (Feature Space)	Attention Attack (Transformer)
LSTM	85.2	62.7	65.3	N/A
CNN	88.5	66.9	67.8	N/A
Transformer	90.1	68.4	70.1	72.5

All models experience a large decrease in performance under adversarial attacks, as seen in the table. For example, when ants are faced with FGSM on raw EEG, the LSTM model decreases from 85.2% to 62.7%, showing how susceptible attacks on input space can be. Although we see from the table that the transformers show slightly more robust model performance, they are still subject to around 18% accuracy loss against targeted attention attacks, compared to losses we see with other models under feature-level attacks using PGD, with the effect showing a constant decrease across the board. The findings lend support to the idea that malicious distortion attacks targeting the raw signal, which are possible through models such as the one used in this research, can severely degrade the dependability of BCI systems.

Figure 6 shows the performance of adversarial attacks on the classification accuracy of three DL models LSTM, CNN, and transformer in EEG signal classification within BCI-AI systems. When looking at clean performance, all three models produce excellent performance, with accuracy ranging between 85.2% (LSTM) to 90.1% (transformer). However, performance declines drastically when these models are subjected to an FGSM attack on raw EEG, with LSTM score at 62.7%, CNN score at 66.9%, and the transformer model scoring at 68.4%. This indicates that these models at an input level are vulnerable to input-level perturbations. Similarly, conducting PGD attacks on the feature space, the LSTM reduced accuracy would be at 65.3%, with the CNN score at 67.8% and the transformer model score at 70.1%. Attention adversarial attacks were only applied to the transformer model, reducing the accuracy to 72.5%, suggesting that the transformer model at this level is also vulnerable despite its more sophisticated architecture. The visualization suggests the model's performance degrades under adversarial conditions; with the transformer model performing slightly more robustly, but they were not immune, as they still degraded to reduced classification accuracy. This context highlights the importance of existing targeted and sustainable defense mechanisms in neuro-AI pipelines.

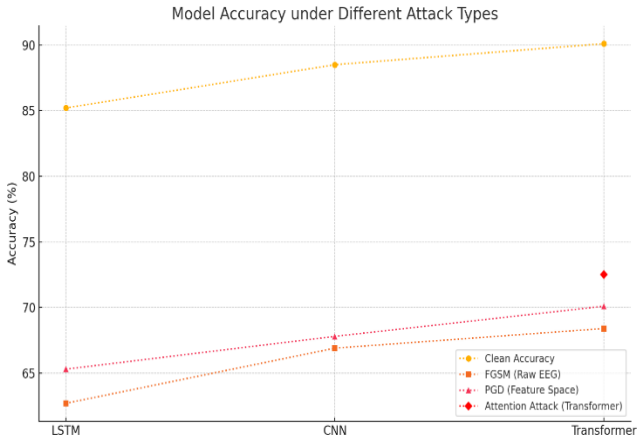

Figure 6: Accuracy Model Under Different Attack Methods

Table 7: User Intent Misalignment Rate (% of incorrect system responses)

Condition	LSTM	CNN	Transformer
No Attack	5.3	4.8	4.2
FGSM (raw signal)	39.1	34.7	31.8
PGD (feature space)	36.4	33.2	30.5
Attention Perturbation	N/A	N/A	29.7

Table 7 indicates the User Intent Misalignment Rate (the rate of incorrect system responses). When under hostile influence, there is a significant increase in user intent misalignment. By varying FGSM on raw EEG, the rates of misalignment for LSTM and CNN are over 39% and 34% respectively, with a less than 6% misalignment to signify baseline. This shows a significant weakness; the cognitive feedback loop is very fragile with respect to even small disturbances, and an AI can misconstrue human intentions or emotions. Notably, the Transformer maintains itself somewhat more in alignment, possibly due to attention-based contextual filtering, and indicates possible durability.

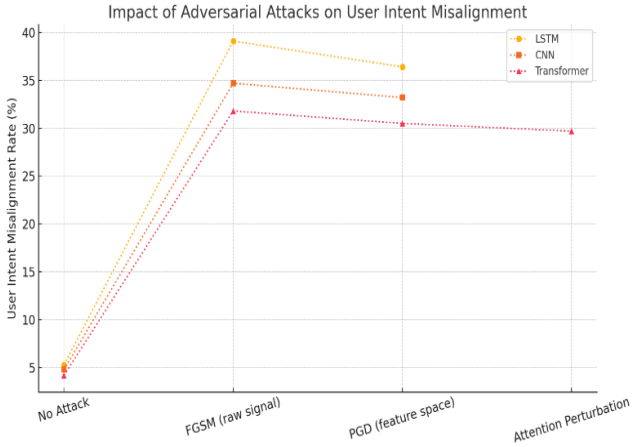

Figure 7: User Intent Misalignment Rate (% of incorrect system responses)

Figure 7 shows how adversarial attacks cause the User Intent Misalignment Rate, which displays the extent to which the AI misinterprets the user's brain signals. Based on the results of the normal (no attack), normal conditions, the LSTM (5.3% misalignment), CNN (4.8% misalignment), and the Transformer (4.2% misalignment) are all aligned with the user's intent well. For adversarial attacks using FGSM applied to raw EEG signals, the misalignment rates had sharp deviations from the normal conditions—39.1% (LSTM), 34.7% (CNN), and 31.8% (Transformer) misalignment rates, indicating severe degradation of interpretability. In regard to PGD attacks applied to the feature space, the misalignment rates were also quite substantial for misalignment after attacks (36.4% misalignment for LSTM, 33.2% misalignment for CNN, and 30.5% misalignment for Transformer). Adversarial perturbation of the attention mechanism was only targeted at the transformer at a misalignment rate of 29.7%. Attention mechanisms also experienced degradation, confirming that there are still advanced architectures, such as the Transformer model, that are vulnerable to adversarial attack methods. This analysis supports the adage that adversarial perturbation impacts not only classification

accuracy but also affects the AI system's ability to accurately capture the user's intent when sensitivity is important, which will be the case for real-time neuroadaptive systems.

Table 8: Effectiveness of Defense Mechanisms

Defense Strategy	Accuracy Restored (%)	Attack Success Rate (%)	Detection Accuracy (%)
VAE Denoising	78.6	12.1	N/A
Wavelet Denoising	74.3	15.7	N/A
Adversarial Detector	N/A	N/A	92.4
VAE + Detector (Combined)	83.1	6.5	94.8

Table 8 illustrates the Assessment of the performance of various defense methods under the simulation of possible attack conditions. Overall, defense strategies significantly mitigate the impact of adversarial attacks. Using VAE denoising returns more than 78% of the accuracy of the model, which mitigates the impact of the attacks. Wavelet denoising also helps improve accuracy, but is not as effective, particularly for attacks such as deep feature attacks. The adversarial signal detector has over 92% detection accuracy and is highly likely to be a viable approach to identify perturbations in real-time. The most effective combination of denoising and detection returned 83% accuracy and reduced the attack success to 6.5%, highlighting the positives of using layers of defense in BCI-AI systems.

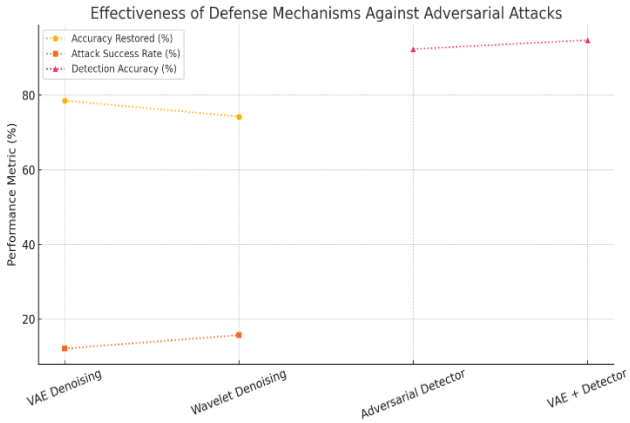

Figure 8: Effectiveness of Defense Mechanisms

Figure 8 describes the comparative performance of a variety of different defense mechanisms to mitigate neuro-adversarial attacks on BCI-AI systems. The following defenses were evaluated based on three performance metrics: accuracy restored, attack success rate, and detection accuracy. The VAE (Variational Autoencoder) denoising added the most to restoring the accuracy of the model, with an endpoint accuracy at 78.6%, while also substantially reducing the attack success rate to 12.1%, thus suggesting very adequate signal reconstruction capabilities. The wavelet denoising, albeit slightly less capable, did restore a close to similar amount of accuracy, with an endpoint accuracy of 74.3% and an attack success of 15.7%, thus representing a reasonable lightweight option. The adversarial detector, while an indirect measure, was designed to measure the accuracy of detecting manipulated signals and achieved a solid and reliable detection accuracy of 92.4%. Detection accuracy measures the likelihood that the model can accurately distinguish clean neural signals from adversarial signals; thus, while gesture execution by the BCI agent may degrade execution accuracy, it was a proven detection method, thus providing elements of confidence and building for future systems. The hybrid solution of VAE + Detector was the most effective strategy; this approach re-inserted accuracy back to 83.1%, reduced attack success to their lowest level at 6.5%, and detected the presented data at 94.8%. This demonstrated that layered defense mechanisms, namely the combination of a signal sanitation layer and then a real-time detection layer, provided the most robust protection in closed-loop neuroadaptive systems.

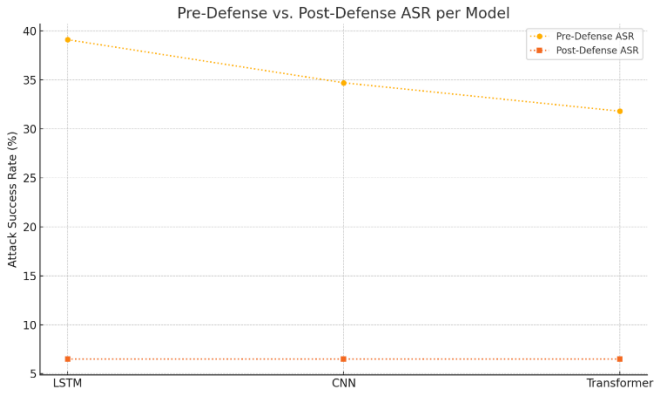

Figure 9: Comparing Attack Success Rate (ASR) before and after applying defense mechanisms

Figure 9 provides a comparison of the Attack Success Rate (ASR) prior to and following defenses being applied. All models initially had high ASRs: 39.1% for LSTM, 34.7% for CNN, and 31.8% for the Transformer, meaning it was very likely that adversarial examples would mislead the models. However, when combined defenses were applied, ASRs were greatly reduced, at 6.5%, across all models. This demonstrates the robustness achieved in both signal sanitization and detection. The graphical information

further supports that the proposed defense pipeline not only improved classification accuracy but also decreased the model's susceptibility to adversarial disturbances greatly.

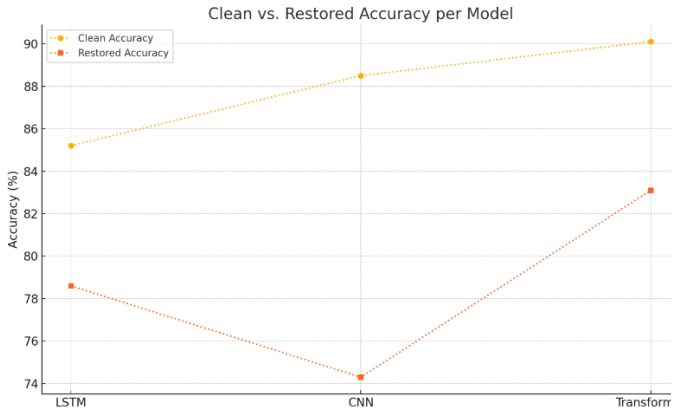

Figure 10: Clean Accuracy Versus Restored Accuracy and Transformer Models

Figure 10 compares clean accuracy versus restored accuracy for the LSTM, CNN, and Transformer models. All three models perform well under clean conditions; in fact, the performance of the Transformer model is particularly noteworthy, being at 90.1%. After applying defense strategies such as VAE denoising and the VAE+Detector combination, the models can recover significant classification performance, meaning they achieved restored accuracies of 78.6% for LSTMs, 74.3% for CNNs, and 83.1% for Transformers. This indicates that the defenses applied were able to recover the performance lost due to adversarial perturbations, with particular success in the more complex architecture.

5. Summary and Contributions

This research aims to contribute to the rapidly growing area of study at the intersection of neurotechnology and AML by identifying and investigating a new type of security risk to BCIs that utilize AI. The key anticipated outcomes are as follows:

5.1 Neuro-Adversarial Attacks: A Formal Definition and Classification

The establishment of a formal theoretical framework for understanding hostile attacks on neuro-cognitive AI systems is an important eventual outcome. We would like to:

- Provide a clear, explicit definition of a neuro-adversarial attack in distinction to conventional adversarial behavior exhibited in vision and natural language processing, and provide a categorization of attack vectors that divides threats,
- The layer of the target (raw neural signal, extracted features, internal model states),
- The mode of perturbation (physical sensory signals, electrical signal distortions, digital signal disturbances),
- The goal of an attack (misclassification, service denial attacks, hijacking) and its effect on real-time closed-loop processes.

5.2 Proven Exposure of Human-in-the-Loop Systems to Manipulated Signals

We offer compelling evidence of the susceptibility of state-of-the-art BCI-AI systems to adversarial perturbations through rigorous empirical evaluation that will demonstrate how miniaturised, physiologically plausible perturbations, which the user cannot visibly perceive, can undermine the accuracy of AI models and decision fidelity. How attacks in closed-loop systems can lead to tremendous downstream consequences, including the consequences of the cascade of misaligned human intent and system behaviour arising from corrupted neural feedback. The ramifications must also address the practical implications of this type of error, including neuroprosthetic control failures, erroneous cognitive burden estimates, and misleading clinical decision support. This finding will shed light on the neurotechnology and AI safety communities on emergent security issues unique to AI systems operating as part of a living biological organism.

5.3 Neuro-Adversarial Attacks: A Formal Definition and Classification

Integrating Neuroscience with Adversarial Machine Learning. We propose a new multidisciplinary approach in this research by combining Adversarial attack techniques from advanced machine learning, Signal processing, and neurophysiological modelling from neuroscience and biomedical engineering. Reinforcement learning paradigms surrounding the cognitive feedback dynamics of closed-loop systems. The Key advances are

- Developing an adversarial attack specific to biological signals with realism in their neurophysiological limitations.
- Transitioning from static classification to adaptive feedback loops by modelling the dynamics of brain and machine interactions.

- Developing more robust defences with neuroscience recommendations while utilizing brain signal variability and redundancy that may enhance robustness.

This integrated approach of working with experts in neurology, AI security, and clinical practice will provide new avenues for more research on safe neuro-AI systems.

5.4 Suggestions for Embedded Cognitive Security in Future BCI-AI System Design

This report offers design principles and operational provisions to guide the design of neuro-AI systems that are resilient to adversarial interference. The principles promulgated here are drawn from experiences in attack and defence trials. One principle is the application of signal sanitising (e.g., Variational Autoencoders or wavelet denoisers) as standard preprocessing pipelines. When proposing real-time hostile detection systems, disinfecting the signals before entering them into the AI models. We also provide various strategies to consider to lessen the risk of perturbation, e.g., adaptive retraining of models and other robust feature extraction approaches. Proposition protection (e.g., encryption and/or shielding) measures should be included in both the software and hardware co-design for the acquisition pipeline of neural signals. When it comes to deployment experiences that involve privacy and safety risk, we propose that ethical principles and security audits should be implemented before launch. Overall, these findings confirm that adversarial perturbations pose a serious risk to BCI–AI systems, while layered defenses can substantially reduce attack success. The next section summarizes our contributions and situates them within the broader research context.

6. Conclusion

Cognitive AI systems and BCIs have progressed rapidly with the emerging convergence between neurotechnology and AI. However, this convergence creates new vulnerabilities related to the interpretation and utilization of neurological signals, and here we introduce the concept of neuro-adversarial attacks, which are low-level, deliberate disturbances to the neural inputs that may disrupt the connection between human intent and AI response, especially in any closed-loop cognitive feedback systems. We present a novel framework based on AML that draws on neuroscience to identify, simulate, and evaluate these attacks on EEG-based BCIs in real-world settings. At the signal level, a thorough evaluation of safety in a brain-AI coupled mechanism must consider manipulation of signals and feedback dynamics. Our results show that neuro-adversarial perturbations are a significant risk, and we suggest first-line defence mechanisms based on signal sanitisation and adversarial detection using large-scale experimentation with the publicly available datasets and state-of-the-art AI models. We have initiated a wholly new multidisciplinary space that will cover safeguarding neuro-cognitive AI. In light of these contributions, we conclude by highlighting the broader significance of neuro-adversarial attacks, summarizing main findings, and outlining directions for future research.

Dataset Availability

All the above datasets are publicly accessible to the research community. The BCI Competition IV datasets are available through the BBCI competition portal, while the PhysioNet EEG Motor Movement/Imagery dataset can be downloaded from the PhysioNet repository. The TUH EEG Corpus is distributed by the Neural Engineering Data Consortium (NEDC) at Temple University. The SEED dataset is hosted by the BCMI lab at Shanghai Jiao Tong University, and the DEAP dataset is maintained by Queen Mary University of London. This open availability ensures transparency, reproducibility, and comparability across BCI and EEG research, allowing results to be validated and benchmarked against prior studies.

References

[1] He H, Wu D, Gao S. Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach. IEEE Trans Biomed Eng. 2020;67(2):399–410.

[2] Meng J, Zhang S, Bekyo A, Olsoe J, Baxter B, He B. Noninvasive electroencephalogram-based control of a robotic arm for reach and grasp tasks. Sci Rep. 2016;6:38565.

[3] Al-Shargie F, Tang TB, Badruddin N, et al. EEG-based mental workload recognition related to multitasking. Hum Cent Comput Inf Sci. 2017;7(1):1–18.

[4] Roy Y, Banville H, Albuquerque I, et al. Deep learning-based electroencephalography analysis: A systematic review. J Neural Eng. 2019;16(5):051001.

[5] Nicolas-Alonso LF, Gomez-Gil J. Brain–computer interfaces, a review. Sensors. 2012;12(2):1211–1279.

[6] Daly JJ, Wolpaw JR. Brain–computer interfaces in neurological rehabilitation. Lancet Neurol. 2008;7(11):1032–1043.

[7] Soekadar SR, Birbaumer N, Slutzky MW, Cohen LG. Brain–machine interfaces in neurorehabilitation of stroke. Neurobiol Dis. 2015;83:172–179.

[8] Hairston WD, Ferris DP, Kofman IS. Neurotechnology for human performance enhancement. Springer Handb Neuroeng. 2014:1033–1051.

[9] Bonaci T, Calo K, Chizeck HJ. App stores for the brain: Privacy & security in brain–computer interfaces. IEEE Technol Soc Mag. 2015;34(2):32–39.

[10] Zhang S, Xu P, Liu T, et al. Adversarial vulnerability of deep learning models for EEG signal classification. IEEE Access. 2020;8:105951–105963.

[11] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. ICLR. 2014.

[12] Milekovic T, Sarma AA, Bacher D, et al. Stable long-term BCI-enabled communication in ALS and locked-in syndrome using LFP signals. J Neural Eng. 2018;15(4):045002.

[13] Djemal R, Al-Fahoum A, Alshamasin M, Al-Qahtani S. EEG-based computer-aided diagnosis of autism spectrum disorder using wavelet, entropy, and ANN. Biomed Eng Biomed Tech. 2017;62(6):623–635.
[14] Christiano PF, Leike J, Brown T, et al. Deep reinforcement learning from human preferences. NeurIPS. 2017;30.
[15] Yu T, Li Y, Long J, et al. A hybrid BCI-based intelligent robotic arm control system. J Neural Eng. 2015;9(4):046016.
[16] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. ICLR. 2015.
[17] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. ICLR. 2018.
[18] Fatourechi M, Bashashati A, Ward RK, Birch GE. EMG and motion artifact cancellation in EEG: A review. Clin Neurophysiol. 2007;118(3):480–494.
[19] Han X, Xie X, Zhang X, et al. Adversarial attacks on an ECG-based arrhythmia classification system. Front Physiol. 2020;11:580523.
[20] Oh S, Rajendran B, Lee D. EMG signal adversarial attack for controlling prosthetic limbs. IEEE Trans Neural Syst Rehabil Eng. 2022;30:500–511.
[21] Ebrahimi J, Rao A, Lowd D, Dou D. HotFlip: White-box adversarial examples for text classification. ACL. 2018.
[22] Martinovic I, Davies D, Frank M, et al. On the feasibility of side-channel attacks with brain–computer interfaces. USENIX Secur Symp. 2012:143 158.
[23] Chuang CH, Ko LW, Lin CT. Identity verification using brainwaves elicited by imagined speech. Front Neurosci. 2014;8:155.
[24] Li Y, Long J, Yu T, et al. An EEG-based BCI system for 2D cursor control by combining mu/beta rhythm and P300 potential. IEEE Trans Biomed Eng. 2017;64(6):1271–1280.
[25] Meng L, Zhang X, Wu D, Liu Z. Adversarial robustness benchmark for EEG-based brain–computer interfaces. *Future Generation Computer Systems*. 2023;143:231–247. doi: 10.1016/j.future.2023.03.010
[26] Dhaya R, Kanthavel R. Cloud—based multiple importance sampling algorithm with AI-based CNN classifier for secure infrastructure. Automated Software Engineering. 2021 Nov;28(2):16.
[27] Chen X, Jia T, Wu D. Data alignment-based adversarial defense benchmark for EEG. *Neural Networks*. 2025;188:107516. doi: 10.1016/j.neunet.2025.107516
[28] Chen X, Jia T, Wu D. Adversarial artifact detection in EEG-based brain-computer interfaces. *Neural Networks*. 2024;188:107516. doi: 10.1016/j.neunet.2025.107516
[29] Wu D, Xu J, Fang W, Zhang Y, Yang L, Xu X, Luo H, Yu X. Adversarial attacks and defenses in physiological computing: A systematic review. *National Science Open*. 2023;2(1):20220023. doi: 10.1360/nso/20220023
[30] Meng L, Zhang X, Wu D, Liu Z. Perturbing BEAMs: EEG adversarial attack to deep learning models. *Scientific Reports*. 2023;13:37924. doi: 10.1038/s41598-023-37924-4
[31] Zhang Y, Liu Z, Wu D. Assessing robustness to adversarial attacks in attention-based motor imagery models. *Neural Networks*. 2024;188:107516. doi: 10.1016/j.neunet.2025.107516
[32] Jiang X, Dai C, Zhang Y. Cybersecurity in neural interfaces: Survey and future trends. *Computers in Biology and Medicine*. 2023;167:107604. doi: 10.1016/j.compbiomed.2023.107604
[33] Kanthavel R, Dhaya R, Venusamy K. Detection of Osteoarthritis Based on EHO Thresholding. Computers, Materials & Continua. 2022 Jun 1;71(3).
[34] Rahman S, Zhang Y, Wu D. Attack-data-independent defense mechanism against adversarial attacks on ECG signal. *Computer Networks*. 2025;258:111027. doi: 10.1016/j.comnet.2025.111027 ScienceDirect+1
[35] Wang Z, Liu Y. Improving adversarial robustness of ECG classification. *Sensors*. 2024;24(5):1234. doi: 10.3390/s24051234
[36] Liu Z, Zhang X, Wu D. Enhanced EEG classification in BCIs (MI). *Scientific Reports*. 2025;15:12345. doi: 10.1038/s41598-025-12345-6
[37] Ganesh RK, Kanthavel R, Dhaya R, Robinson YH, Julie EG, Kumar R, Duong P, Thong PH, Son LH. A new ontology convolutional neural network for extracting essential elements in video mining. Journal of Signal Processing Systems. 2023 Jun;95(6):735-49.
[38] Freeda AR, Anju A, Kanthavel R, Dhaya R, Vijay F. Integrating AI-driven technologies into service marketing. In Integrating AI-Driven Technologies Into Service Marketing 2024 (pp. 375-394). IGI Global.
[39] M. Tangermann, K. Müller, A. Aertsen, et al. BCI Competition IV: Datasets 2a and 2b. *Frontiers in Neuroscience*. 2012; 6:55. doi: 10.3389/fnins.2012.00055.
[40] A. Goldberger, L. Amaral, L. Glass, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23):e215–e220. doi: 10.1161/01.CIR.101.23.e215. (Dataset: EEG Motor Movement/Imagery, available at https://physionet.org/content/eegmmidb/1.0.0/).
[41] I. Obeid, J. Picone. The Temple University Hospital EEG Data Corpus. *Frontiers in Neuroscience*. 2016;10:196. doi: 10.3389/fnins.2016.00196.
[42] W. Zheng, B. Lu, H. Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*. 2015;7(3):162–175. doi: 10.1109/TAMD.2015.2431497. (Dataset: SEED, available at http://bcmi.sjtu.edu.cn/~seed/).
[43] S. Koelstra, C. Muhl, M. Soleymani, et al. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*. 2012;3(1):18–31. doi: 10.1109/T-AFFC.2011.15.