A First Look at the Security of EEG-based Systems and Intelligent Algorithms under Physical Signal Injections

Md Imran Hossen*
md-imran.hossen1@louisiana.edu
University of Louisiana at Lafayette
Lafayette, Louisiana, USA

Yazhou Tu*
yazhou.tu1@louisiana.edu
University of Louisiana at Lafayette
Lafayette, Louisiana, USA

Xiali Hei
xiali.hei@louisiana.edu
University of Louisiana at Lafayette
Lafayette, Louisiana, USA

ABSTRACT

Electroencephalography (EEG) based systems utilize machine learning (ML) and deep learning (DL) models in various applications such as seizure detection, emotion recognition, cognitive workload estimation, and brain-computer interface (BCI). However, the security and robustness of such intelligent systems under analog-domain threats have received limited attention. This paper presents the first demonstration of physical signal injection attacks on ML and DL models utilizing EEG data. We investigate how an adversary can degrade the performance of different models by non-invasively injecting signals into EEG recordings. We show that the attacks can mislead or manipulate the models and diminish the reliability of EEG-based systems. Overall, this research sheds light on the need for more trustworthy physiological-signal-based intelligent systems in the healthcare field and opens up avenues for future work.

CCS CONCEPTS

• **Security and privacy → Embedded systems security**.

KEYWORDS

EEG, neural network, machine learning, epilepsy diagnosis, EMI

ACM Reference Format:
Md Imran Hossen, Yazhou Tu, and Xiali Hei. 2023. A First Look at the Security of EEG-based Systems and Intelligent Algorithms under Physical Signal Injections. In *Proceedings of the 2023 Secure and Trustworthy Deep Learning Systems Workshop (SecTL '23), July 10–14, 2023, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3591197.3591304

1 INTRODUCTION

The electroencephalography (EEG) signal is a physiological signal widely used in neuroscience research and clinical settings for diagnosing and monitoring various neurological disorders. In recent years, machine learning (ML) and deep learning (DL) have become increasingly important in various EEG applications, including but not limited to seizure detection in epilepsy patients [33, 34], emotion recognition [43], cognitive workload estimation [37], and brain-computer interface (BCI) applications [7, 10].

While EEG has many other potential applications in neuroscience, psychology research, and BCIs, its use in epilepsy diagnosis and treatment is of particular importance due to the potentially life-threatening nature of the condition [16, 17]. ML and DL-based seizure detection models have recently gained significant attention as promising tools for diagnosing epilepsy, with the ability to accurately and automatically identify seizures from EEG signals.

*Both authors contributed equally to this research.

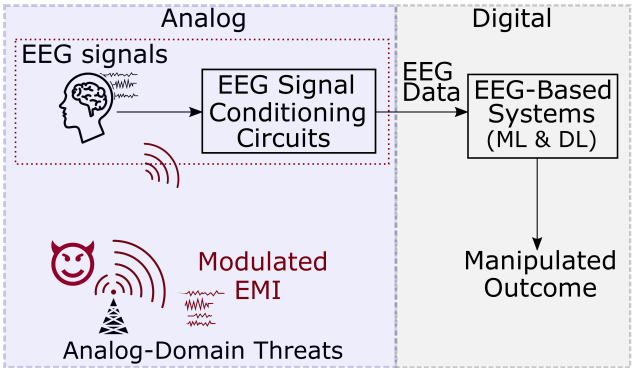

Figure 1: Our methodology is based on non-invasive signal injections on EEG-based systems. Adversaries can remotely modulate out-of-band electromagnetic signals to inject fabricated brain wave signals into the circuits.

Despite their potential to greatly improve diagnosis outcomes, the security and robustness of these systems under physical signal injection attacks are yet to be investigated. Such attacks pose a serious threat to machine learning-based systems in healthcare applications, and the consequences can be severe. For instance, in the case of seizure detection systems, the physical signal injection attack could cause the system to misclassify normal brain activity as a seizure or fail to detect a genuine seizure, leading to inappropriate or delayed interventions that could result in adverse outcomes such as brain damage, injury, or even death [25, 26].

This paper explores analog-domain threats to EEG-based systems. Specifically, we study the security and robustness of EEG-based systems under electromagnetic interference (EMI) signal injections (Fig. 1). We show that, by modulating electromagnetic signals, external adversaries can non-invasively inject data into EEG-based systems without acquiring permission to directly tamper with the dataset or the models.

We further identify the attack surface of EEG-based systems under analog-domain attacks and characterize the attack mechanisms. We find that different frequencies and waveforms of modulated EMI signals can be remotely injected into the EEG data perceived by the system, allowing adversaries to design specific types of attacks to exploit EEG-based intelligent systems.

Our pilot study demonstrates the emerging threat of non-invasive attacks on EEG-based systems. Our work shows that signal injection attacks can intentionally manipulate the results or significantly degrade the performance of the models. Prior works focused on

digital-domain threats exploiting the data, software, and communication of brain-computer interfaces [6, 24]. Our results highlight the vulnerabilities of data acquisition, learning, and processing in EEG-based systems under analog signal injections.

In summary, this paper makes the following contributions:

- We develop the first study on the security and robustness of EEG-based systems under physical signal injections.
- We observe that different modulated EMI signals can be non-intrusively injected into EEG-based systems, allowing adversaries to design attacks to mislead or intentionally manipulate the EEG-based ML and DL models without requiring access to the original data.
- We evaluate our attacks on different ML/DL models. The attacks could cause the system to perceive normal brain activity as a seizure or fail to detect a genuine seizure. Furthermore, it remains challenging to filter the attack signals. Our study highlights the need for more trustworthy EEG-based data acquisition, processing, and intelligent systems.

2 BACKGROUND

EEG Signals. Brain cells communicate via electrical impulses. Electroencephalography (EEG) signals are physiological signals generated from brain activities. Depending on the technology, electrodes may be placed on the scalp or in the substance of the patient [29] to collect EEG signals that are of the order of microvolts [19]. The processing of EEG signals usually requires bandpass filtering and notch filtering to remove noise. Epilepsy is a chronic, noncommunicable brain disease characterized by recurrent epileptic seizures. These seizures occur due to temporary excessive electrical activity in the brain. Such abnormal and excessive electrical discharging activities can be recorded and processed using EEG-based systems for epileptic diagnosis.

EEG-based Models. This paper mainly focuses on machine learning (ML) and deep learning (DL) models that utilize EEG signals for seizure detection. Developing EEG-based ML systems for seizure detection typically involves several key steps. Firstly, the EEG data must be preprocessed to remove noise and artifacts, which can be done using various signal processing techniques, such as filtering and artifact removal algorithms. Next, relevant features are extracted from the preprocessed EEG signals. These features could include statistical, time domain, frequency domain, time-frequency domain, and nonlinear parameters, depending on the requirements of specific tasks [2, 32]. Once the features are extracted, a classifier is trained on labeled data to identify EEG data patterns indicative of seizures. Various types of classifiers have been used for this task, including support vector machines (SVMs), decision trees, and random forest [34].

Recently, DL techniques have become popular for EEG-based seizure detection tasks due to their ability to automate feature extraction, perform end-to-end learning, and achieve state-of-the-art results. Unlike traditional machine learning techniques, which often require manual feature engineering and selection, DL models can learn complex features directly from raw EEG data. In recent years, various DL architectures have been proposed for seizure detection, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their combinations, such as hybrid CNN-RNN models [33]. These models have shown promising results and outperformed traditional ML techniques on various benchmark datasets.

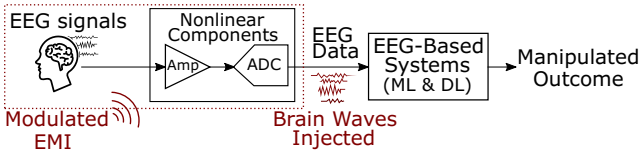

Figure 2: In physical brain wave injection attacks, the out-of-band EMI signals will be nonlinearly transformed in the circuits, allowing the fabricated brain wave signals to affect EEG-based ML and DL systems.

3 METHODOLOGY

This section investigates how adversaries can attack EEG-based intelligent systems in a non-invasive setting. DL and ML models are vulnerable to attacks that directly tamper with the samples in the digital domain. However, such attacks rely on strong assumptions about the attacker's capability and permission to directly modify the internal training and testing data. The attackers also require prior knowledge about the model structure and parameters of the victim system, which is commonly referred to as a white-box attack. This work addresses this issue by considering attacks that do not require explicit modification or access to the victim system's internal data or model. Furthermore, we investigate the following question: can adversaries inject signals with components similar to brainwave signals [36] to manipulate DL and ML-based systems?

3.1 Brain Wave Injection

We first inject simple-pattern signals with different frequencies. We then inject signal components in specific brain wave bands [36].

Injection of simple signals. We modulate the adversarial signal $a(t)$ onto a high-frequency out-of-band electromagnetic carrier. The carrier signal enters the analog circuits of the EEG-based systems via electromagnetic coupling. After demodulating [31] the interfering out-of-band signals in the analog circuit, the adversarial signal $a'(t)$ will be superimposed [31] onto the original EEG signal and influence the EEG-based systems. We illustrate this process in Fig. 2. EEG acquisition devices are equipped with amplifiers and analog-to-digital converters (ADCs) to measure EEG signals in the order of microvolts [40]. Nonlinear sensor components (e.g., microphones [11, 12, 18, 20, 30, 42]), including amplifiers and ADCs, can allow for the demodulation of EMI signals [31].

We inject modulated EMI signals into an EEG electrode cap connected to the OpenBCI Cyton biosensing board using an antenna, as explained in Section 4.1. We make observations on the injected data in both the time and frequency domains.

Observations. We observe that adversaries can inject sine wave signals of different frequencies into the system using amplitude modulation. The EMI signals will be demodulated by the non-linear components in the analog circuits of EEG devices [31]. We find that the injected sine wave signals are usually not ideal because

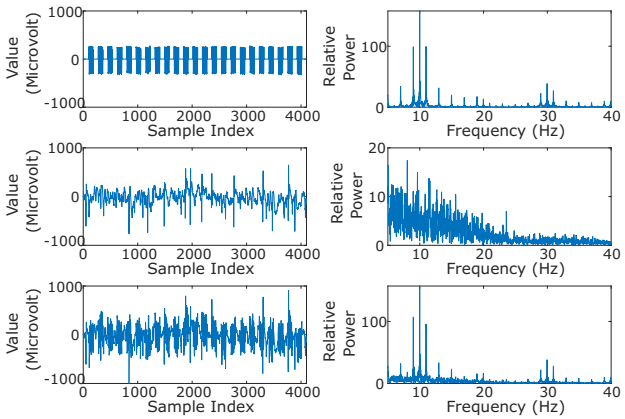

Figure 3: Injecting modulated EMI signals. We inject 10-Hz sine wave signals with 0.5-s intervals. Top: Injection data. Middle: Original EEG data. Bottom: EEG data under injection.

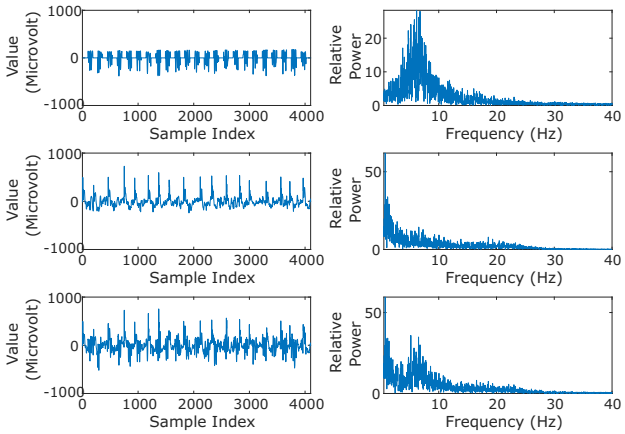

Figure 4: Injecting modulated brain wave (Theta wave) band noise to original data. Top: Injeciton data. Middle: Original EEG data. Bottom: EEG data under injection.

the demodulation of the out-of-band EMI signals in analog circuits is usually imperfect. For example, when we modulate a 10-Hz sine wave signal with high-frequency (573 MHz) electromagnetic carrier waves, the demodulated signal in the EEG system shows that the demodulation can generate artifacts that include other frequency components (see Fig. 3, top right).

Although the injected simple-pattern signals, such as sine and square waves, can affect the outcome of EEG-based systems (see Fig. 6 and Tab. 2), they may be easily detected or filtered. We can observe that the injected signals can become evident in the frequency domain (see Fig. 3, bottom right), even though they are not significant in the time domain.

Injection of brain wave components. Inspired by the physics that human brain wave signals usually reside in bands categorized as Theta, Alpha, Mu, Beta, etc. [36], we explore the effects of more targeted injections of selective brain-wave-band components on EEG-based systems. To inject such brain-wave-band components, we process white noise signals with Butterworth band-pass filters

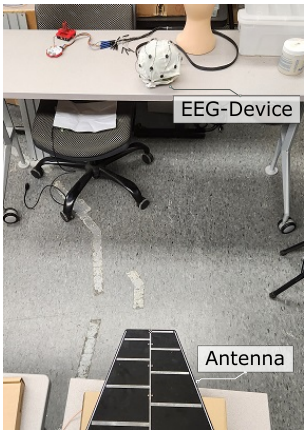

Figure 5: Injection Settings. We use an antenna to emit the modulated EMI signals.

that correspond to the frequency ranges of specific brain wave bands. We then modulate the band-pass filtered noise signals with 573-MHz carrier electromagnetic waves.

Observations. Compared to injecting simple-pattern signals such as sine wave signals, the injected brain wave band noise can have a smaller magnitude but affect the performance of the models more significantly (Fig. 7). This is because the time and frequency-domain features in specific brain wave bands can be affected by the injection. Such attacks can be particularly challenging for victim systems to mitigate because the injected brain-wave-band components are blended with the original EEG signals in both time and frequency domains (see Fig. 4). Additionally, brain-wave-band noise lacks a distinct pattern like the sine wave, making it difficult to detect. The EEG-based systems will perceive the injected signal as genuine EEG signals, resulting in false outcomes for different models (see Section 4.3).

3.2 Threat Model

The adversary can use antennas to non-invasively inject signals from a range of one to several meters. They may also increase the transmitting power and utilize directional antennas to inject signals from a longer distance. Capable adversaries may even generate electromagnetic attack signals using signal towers or radars. The adversaries could also use portable EMI-emitting devices. For example, they may use off-the-shelf software-defined radio (SDR) devices [1] to create a portable attack device that can be carried in a backpack. The attack can be launched in non-line-of-sight settings since EMI signals can penetrate many materials. The attack device could be hidden under, inside, or behind a table, box, wall, or other objects. We assume that the adversaries can induce signals in a normal range in the analog circuits but usually cannot inject signals that are much stronger than seizure signals or damage the components.

We assume that adversaries may not directly modify the dataset. Moreover, we consider a black-box attack setting where the adversary does not have access to or knowledge of the internal structure or parameters of the target or victim models.

Table 1: Performance of models on the clean test set.

Model	Accuracy	Precision	Recall	F1
LR	0.9920	0.9923	0.9920	0.9921
SVM	0.9920	0.9923	0.9920	0.9921
DT	0.9760	0.9765	0.9760	0.9762
RF	0.9920	0.9923	0.9920	0.9921
KNN	0.9680	0.9678	0.9680	0.9675
ConvNet1D	0.9920	0.9923	0.9920	0.9921

4 EXPERIMENTAL EVALUATION

4.1 Settings

Physical signal injection. As shown in Fig. 5, we place the electrodes on the surface of a watermelon, which serves as a conductive head phantom. This setting is similar to that in [5]. We use the watermelon because its surface is conductive, like the human scalp, thus providing a safe and inexpensive means to investigate the risks of physical signal injections on EEG-based systems. We inject physical signals with a directional antenna from a distance of 1.04 meter.

Data. Our study uses the Bonn University EEG dataset [4] for baseline evaluation. The dataset has been used in numerous studies to develop and evaluate algorithms for classifying EEG signals into normal and seizure categories. The dataset comprises 500 EEG segments, each lasting 23.6 seconds. The EEG signals were filtered using a bandpass filter of 0.53-40 Hz. 100 EEG samples represent seizure activity, and 400 samples are considered as normal. In our attack evaluation, we scale the magnitude of EMI injected signals collected from our device to comparable ranges of the samples from the Bonn University dataset.

4.2 Baseline

Numerous machine learning methods use statistical, time, frequency, time-frequency domain, and nonlinear parameters to detect epileptic seizures [34]. In this study, we extract various standard frequency and time-frequency domain features, including power spectral density (PSD) and wavelet coefficients, using the Welch and Discrete Wavelet Transform (DWT) techniques. After extracting features from the preprocessed EEG data, we use them to train machine learning algorithms for seizure detection. Specifically, we use five traditional ML models: logistic regression (LR), support vector machine (SVM), decision tree (DT), and k-nearest neighbors (KNN). We divide the data into training and testing sets in a 75:25 ratio, with 75% used for training and the remaining 25% used for testing. We use a 5-fold cross-validation strategy to optimize hyperparameters and prevent overfitting. Table 3 (see Appendix C) summarizes the process of exploring hyperparameters and presents the optimal hyperparameters obtained for each model through grid search. For each model, hyperparameters were explored within a predefined range of values, and the optimal hyperparameters were selected based on the weighted F1 score.

We also use a 1D convolutional neural network (ConvNet1D)-based DL model consisting of five Convolution-ReLU-MaxPool layers and two fully connected layers. The model is trained on pre-processed EEG data without explicit feature engineering, using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. To prevent overfitting, we employ the early stopping technique, monitoring the model's performance on the validation set and stopping training once the performance stops improving or begins to degrade.

Table 2: Performance of different models under physical brain wave injection attacks using different modulated signals.

Injection	Model	Accuracy	Precision	Recall	F1
Sine wave (10 Hz)	LR	0.4720	0.8549	0.4720	0.4922
	SVM	0.2880	0.8439	0.2880	0.2305
	DT	0.8800	0.9030	0.8800	0.8863
	RF	0.8320	0.9087	0.8320	0.8470
	KNN	0.7920	0.8874	0.7920	0.8115
	ConvNet1D	0.3680	0.8481	0.3680	0.3552
Square wave (10 Hz)	LR	0.5040	0.8575	0.5040	0.5299
	SVM	0.2400	0.8417	0.2400	0.1452
	DT	0.8160	0.8758	0.8160	0.8308
	RF	0.7600	0.8909	0.7600	0.7838
	KNN	0.6800	0.8769	0.6800	0.7111
	ConvNet1D	0.2560	0.8424	0.2560	0.1746
Brain-wave-band noise (Theta wave band 4-7 Hz)	LR	0.6720	0.8758	0.6720	0.7036
	SVM	0.2720	0.8431	0.2720	0.2030
	DT	0.6160	0.8685	0.6160	0.6494
	RF	0.2000	0.0400	0.2000	0.0667
	KNN	0.2720	0.8431	0.2720	0.2030
	ConvNet1D	0.2000	0.0400	0.2000	0.0667

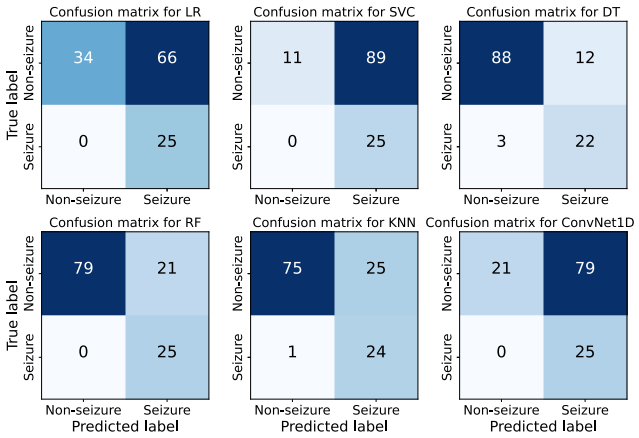

Figure 6: Confusion matrices for different models on the injected test set (sine wave, 10 Hz).

Table 1 depicts the performance results of traditional machine learning algorithms and the deep learning model on the clean test set. We use standard performance metrics like accuracy and F1-score to discuss the results (see Appendix B). The confusion matrices in Fig. 8 (see Appendix D) visually represent the models' predictions. All ML models achieve high accuracy, with F1 scores ranging from 96% to 99%. The LR, SVM, and RF classifiers obtain the highest F1 scores, all above 99%, while DT and KNN perform slightly worse but still achieve F1 scores above 96%. The CovNet1D DL model also achieves an F1 score greater than 99%.

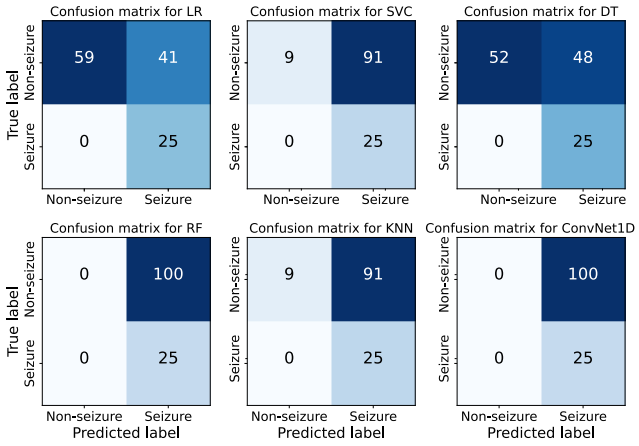

Figure 7: Confusion matrices for different models with the EMI injected Theta-wave noises.

4.3 Attack Evaluation

We inject different types of signals, including sine wave, square wave, and brain-wave-band noise, with varying frequencies into the EEG device to assess the effectiveness of our attack. We then add the scaled EMI-injected signals to the original samples in the test set to simulate a real attack and gather the models' predictions. We filter all samples using a bandpass filter of 0.53–40 Hz before feeding them to the classifiers as inputs. Table 2 shows the performance of the models on the perturbed test set, i.e., the EEG signals with injected EMI. We can observe that most of the models' accuracy and F1-score significantly decrease under the physical injection attack. For the 10 Hz sine wave injection, the accuracy of LR, SVM, and ConvNet1D classifiers falls below that of a random guess, rendering them useless. The performances of DT, RF, and KNN are relatively less impacted. We notice similar results for the 10-Hz square wave as well. The most interesting results can be observed for the brain-wave-band noise injection, which significantly degrades the F1-score of all the models. The SVM, RF, KNN, and ConvNet1D models' performances are impacted to the point that they become completely useless for the seizure detection task.

Figure 6 and Figure 7 show the confusion matrices for all the models under physical injection attack with a 10 Hz sine wave and theta wave band noise, respectively. We found that most of the wrongly predicted samples belonged to the non-seizure class, but they were predicted as seizures after the injection attack. We also noticed similar trends for other injected waves in our experiments. Due to space constraints, we have included more results in Appendix D.

5 DISCUSSION

Mitigation. The use of shielding can help reduce certain types of interference in EEG data acquisition systems but may also increase their weight. Shielded sensors can still be susceptible to intentional EMI signal injections [38]. While shielded rooms can effectively reduce some types of noise in EEG recordings, such as power line interference, they may not completely eliminate all sources of interference [21]. Researchers also proposed methods to detect EMI signals or extract the original sensor signals [39, 45]. However, since EEG devices are highly sensitive compared to other sensors, any attack mitigation methods must be carefully tested and verified to ensure that they do not affect the normal functioning and usability of EEG-based systems.

It may be possible to filter specific noises, such as white noises or signals at a specific frequency. However, adversaries can intentionally inject different kinds of signals, such as brain-wave-band noises. As the ML and DL models rely on information extracted from such components, filtering out such signals without significantly impacting the original signals can be challenging. Moreover, EEG signals are highly uncertain and usually contain more noise than known signals, such as ECG [20] and replayed human voices [20, 42], making it more difficult to distinguish the attacks.

Limitations and Future Work. We have observed that during our physical injection attack, most non-seizure events are misclassified as seizure events. It would be intriguing to investigate whether an attacker could create specific waves to induce targeted misclassifications, such as classifying seizure events as non-seizure events or vice versa. The attack distance is related to the transmitting power and the gain [20]. Although the attack distance can be increased with higher transmitting power and specialized electromagnetic transmitters (e.g., high-gain antennas), our goal is to demonstrate that physical signal injections can induce specific brain-wave-band components to mislead EEG-based intelligent systems. In this study, we do not assume that the adversaries can access the user's EEG recordings and replay the signals by modulating it on the EMI carrier waves. We observe that brain-wave-band noise injections induce fewer artifacts than sine wave injections (Section 3); we plan to conduct more in-depth analysis and evaluation of the underlying mechanisms and attack effects in the future.

6 RELATED WORK

Physical injection attacks on physiological signals. Attacks on physiological signals can have critical and even life-threatening implications. Prior works studied ECG signal injections based on in-band attack signals [9, 20]. In-band EMI attacks [20] on cardiac implantable electrical devices (CIEDs) can inhibit pacing and induce defibrillation shocks [20]. In comparison to attacks on ECG signals, our attacks are based on out-of-band signal injections targeting more complex EEG signals. Further, we investigate the threats on the security and robustness of EEG-based intelligent systems.

Other studies related to physiological signal injections targeted skin temperature [38] and blood glucose level [15, 28]. These works primarily focused on manipulating sensor values within the context of medical control systems, such as infant incubators and artificial pancreases, and did not study the threats to systems based on ML or DL models.

Adversarial attacks on machine learning systems. Previous research has shown that machine learning (ML) is vulnerable to adversarial attacks [13, 35]. There are two types of adversarial attacks based on the stage at which they are executed: poisoning attacks and evasion attacks. Both of these attacks have been extensively studied, mostly in the computer vision domain [8, 14, 22, 41]. However, physiological signals are time series that are continuous in

nature and are distinct from images. There have been relatively limited investigations into adversarial attacks targeted at time series data. Our attack falls under the category of evasion attack, and we discuss below a few of such recently published works.

Zhang et al. [44] showed that deep learning models used in EEG-based Brain-Computer Interfaces (BCIs) can be vulnerable to adversarial attacks. They conducted non-targeted evasion attacks on three CNN classifiers in three BCI paradigms using a jamming module and optimized the adversarial examples using unsupervised FGSM. Aminifar [3] conducted a study on EEG-based epileptic seizure detection and explored the use of Universal Adversarial Perturbations (UAPs) in targeted evasion attacks. The approach used was a white-box attack, which assumes complete knowledge of the target classifier. Newaz et al. [27] conducted adversarial attacks in machine learning-based smart healthcare systems consisting of 10 vital signs, such as EEG, ECG, SpO2, respiration, blood pressure, blood glucose, blood hemoglobin, and others. They carried out targeted and non-targeted poisoning and evasion attacks on both white-box and black-box models, and demonstrated that such attacks can considerably reduce the accuracy of four different classifiers in detecting diseases and normal activities, potentially resulting in incorrect treatments.

While the attack strategies discussed above are promising in theory, their practical application faces several challenges. One significant challenge is trial-specificity, which requires the attacker to generate distinct adversarial perturbations for each EEG trial. Additionally, the attacker needs to gain direct access to and modify the benign sample digitally in order to create an adversarial example that will be received by the system. Furthermore, crafting adversarial examples using optimization techniques can be computationally expensive and often requires knowledge of the target model [8, 23]. Adversarial perturbations are designed to fool the targeted classifier and may not be effective against other models. For example, adversarial perturbations exploit specific weaknesses in a targeted classifier's decision-making process that may not exist or be different in other models. Additionally, the effectiveness of perturbations can vary depending on a model's architecture, training data, and optimization algorithm. These challenges limit the practicality and applicability of existing attack methods, making them less effective in real-world scenarios.

In contrast to existing methods, our approach to adversarial attacks on EEG-based intelligent systems is simple, model-agnostic, and can be executed in real-time. The adversaries do not need permission to access or modify the digital EEG samples to launch the attack, as they can corrupt the EEG recordings by non-invasively injecting EMI signals into EEG acquisition systems. Our method is potentially more dangerous, as it can evade a wide range of classifiers without requiring any specific knowledge of the target models. This makes our method highly effective in real-world scenarios where an attacker may not have prior knowledge of the target system.

7 CONCLUSION

We studied non-invasive attacks on EEG-based systems. Our work demonstrated the emerging threats of manipulating critical bioelectric signal-based intelligent systems with physical signal injections. We evaluated our attack against various commonly used machine learning algorithms and a deep learning model for seizure detection. Results show that the physical injection attacks can significantly degrade the performance of these ML/DL models. We also discovered that by injecting specific brain-wave-band noises, adversaries could manipulate the EEG-based intelligent systems without prior knowledge of the model or data. Because brain-wave-band signals are essential for EEG-based applications, it is difficult to eliminate only the injected components. Our work highlights the need for more trustworthy bioelectric-signal-based measuring, processing, and decision-making to improve the safety and reliability of intelligent systems utilizing complex EEG signals.

Acknowledgement. This work is partly supported by the US NSF under grants CNS-1650551, OIA-1946231, CNS-2117785, and OIA-2229752.

REFERENCES

[1] 2022. HackRF One. https://greatscottgadgets.com/hackrf/one/.
[2] Emina Alickovic, Jasmin Kevric, and Abdulhamit Subasi. 2018. Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. *Biomedical signal processing and control* 39 (2018), 94–102.
[3] Amir Aminifar. 2020. Universal adversarial perturbations in epileptic seizure detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–6.
[4] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. 2001. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E* 64, 6 (2001), 061907.
[5] Mukund Balasubramanian, William M Wells, John R Ives, Patrick Britz, Robert V Mulkern, and Darren B Orbach. 2017. RF heating of gold cup and conductive plastic electrodes during simultaneous EEG and MRI. *The Neurodiagnostic Journal* 57, 1 (2017), 69–83.
[6] Sergio López Bernal, Alberto Huertas Celdrán, and Gregorio Martínez Pérez. 2023. Eight Reasons to Prioritize Brain-Computer Interface Cybersecurity. *Commun. ACM* 66, 4 (2023), 68–78.
[7] Zehong Cao. 2020. A review of artificial intelligence for EEG-based brain- computer interfaces and applications. *Brain Science Advances* 6, 3 (2020), 162–170.
[8] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. Ieee, 39–57.
[9] Simon Eberz, Nicola Paoletti, Marc Roeschlin, Marta Kwiatkowska, I Martinovic, and A Patané. 2017. Broken hearted: How to attack ECG biometrics. In *Network and Distributed Systems Security (NDSS) Symposium*.
[10] Bradley J Edelman, Jianjun Meng, Daniel Suma, Claire Zurn, E Nagarajan, BS Baxter, Christopher C Cline, and BJSR He. 2019. Noninvasive neuroimaging enhances continuous neural tracking for robotic device control. *Science robotics* 4, 31 (2019), eaaw6844.
[11] J Lopes Esteves and C Kasmi. 2018. Remote and silent voice command injection on a smartphone through conducted IEMI: Threats of smart IEMI for information security. *Wireless Security Lab, French Network and Information Security Agency (ANSSI), Tech. Rep* (2018).
[12] Ilias Giechaskiel, Youqian Zhang, and Kasper B Rasmussen. 2019. A framework for evaluating security in the presence of signal injection attacks. In *Computer Security–ESORICS 2019: 24th European Symposium on Research in Computer Security, Luxembourg, September 23–27, 2019, Proceedings, Part I 24*. Springer, 512–532.
[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
[14] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733* (2017).
[15] Xiali Hei and Yazhou Tu. 2021. Glucose monitorying method and system. US Patent App. 16/952,692.
[16] John Hewertson, Martin P Samuels, David P Southall, Christian F Poets, Stewart G Boyd, and Brian GR Neville. 1994. Epileptic seizure-induced hypoxemia in infants with apparent life-threatening events. *Pediatrics* 94, 2 (1994), 148–156.
[17] Nathalie Jette and Jerome Engel. 2016. Refractory epilepsy is a life-threatening disease: Lest we forget. , 1932–1933 pages.
[18] Chaouki Kasmi and Jose Lopes Esteves. 2015. IEMI threats for information security: Remote command injection on modern smartphones. *IEEE Transactions on Electromagnetic Compatibility* 57, 6 (2015), 1752–1755.

[19] V Krishnaveni, S Jayaraman, S Aravind, V Hariharasudhan, and K Ramadoss. 2006. Automatic identification and removal of ocular artifacts from EEG using wavelet transform. *Measurement science review* 6, 4 (2006), 45–57.
[20] Denis Foo Kune, John Backes, Shane S Clark, Daniel Kramer, Matthew Reynolds, Kevin Fu, Yongdae Kim, and Wenyuan Xu. 2013. Ghost talk: Mitigating EMI signal injection attacks against analog sensors. In *2013 IEEE Symposium on Security and Privacy*. IEEE, 145–159.
[21] Sabine Leske and Sarang S Dalal. 2019. Reducing power line noise in EEG and MEG data via spectrum interpolation. *Neuroimage* 189 (2019), 763–776.
[22] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2022. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).
[24] Ivan Martinovic, Doug Davies, Mario Frank, Daniele Perito, Tomas Ros, and Dawn Song. 2012. On the Feasibility of {Side-Channel} Attacks with {Brain-Computer} Interfaces. In *21st USENIX Security Symposium (USENIX Security 12)*. 143–158.
[25] CPJA Monté, JBAM Arends, IY Tan, AP Aldenkamp, M Limburg, and MCTFM De Krom. 2007. Sudden unexpected death in epilepsy patients: risk factors: a systematic review. *Seizure* 16, 1 (2007), 1–7.
[26] Maromi Nei, Reginald T Ho, Bassel W Abou-Khalil, Frank W Drislane, Joyce Liporace, Alicia Romeo, and Michael R Sperling. 2004. EEG and ECG in sudden unexplained death in epilepsy. *Epilepsia* 45, 4 (2004), 338–345.
[27] AKM Iqtidar Newaz, Nur Imtiazul Haque, Amit Kumar Sikder, Mohammad Ashiqur Rahman, and A Selcuk Uluagac. 2020. Adversarial attacks to machine learning-based smart healthcare systems. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 1–6.
[28] Md Fazle Rabby, Yazhou Tu, Md Imran Hossen, Insup Lee, Anthony S Maida, and Xiali Hei. 2021. Stacked LSTM based deep recurrent neural network with kalman smoothing for blood glucose prediction. *BMC Medical Informatics and Decision Making* 21 (2021), 1–15.
[29] Georgia Ramantani, Louis Maillard, and Laurent Koessler. 2016. Correlation of invasive EEG and scalp EEG. *Seizure* 41 (2016), 196–200.
[30] Kasper Bonne Rasmussen, Claude Castelluccia, Thomas S Heydt-Benjamin, and Srdjan Capkun. 2009. Proximity-based access control for implantable medical devices. In *Proceedings of the 16th ACM conference on Computer and communications security*. 410–419.
[31] Jean-Michel Redouté and Michiel Steyaert. 2009. *EMC of analog integrated circuits*. Springer Science & Business Media.
[32] Manish Sharma, Ankit A Bhurane, and U Rajendra Acharya. 2018. MMSFL-OWFB: A novel class of orthogonal wavelet filters for epileptic seizure detection. *Knowledge-Based Systems* 160 (2018), 265–277.
[33] Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Mahboobeh Jafari, Parisa Moridian, Roohallah Alizadehsani, Maryam Panahiazar, Fahime Khozeimeh, Assef Zare, Hossein Hosseini-Nejad, et al. 2021. Epileptic seizures detection using deep learning techniques: A review. *International Journal of Environmental Research and Public Health* 18, 11 (2021), 5780.
[34] Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Xiaodi Huang, and Nasir Hussain. 2020. A review of epileptic seizure detection using machine learning classifiers. *Brain informatics* 7, 1 (2020), 1–18.
[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).
[36] Michal Teplan et al. 2002. Fundamentals of EEG measurement. *Measurement science review* 2, 2 (2002), 1–11.
[37] Christoph Tremmel, Christian Herff, Tetsuya Sato, Krzysztof Rechowicz, Yusuke Yamani, and Dean J Krusienski. 2019. Estimating cognitive workload in an interactive virtual reality environment using EEG. *Frontiers in human neuroscience* 13 (2019), 401.
[38] Yazhou Tu, Sara Rampazzi, Bin Hao, Angel Rodriguez, Kevin Fu, and Xiali Hei. 2019. Trick or heat? Manipulating critical temperature-based control systems using rectification attacks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2301–2315.
[39] Yazhou Tu, Vijay Srinivas Tida, Zhongqi Pan, and Xiali Hei. 2021. Transduction shield: A low-complexity method to detect and correct the effects of EMI injection attacks on sensors. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*. 901–915.
[40] Naveen Verma, Ali Shoeb, Jose Bohorquez, Joel Dawson, John Guttag, and Anantha P Chandrakasan. 2010. A micro-power EEG acquisition SoC with integrated feature extraction processor for a chronic seizure detection system. *IEEE journal of solid-state circuits* 45, 4 (2010), 804–816.
[41] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems* 30, 9 (2019), 2805–2824.
[42] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 103–117.
[43] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 2020. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion* 59 (2020), 103–126.
[44] Xiao Zhang and Dongrui Wu. 2019. On the vulnerability of CNN classifiers in EEG-based BCIs. *IEEE transactions on neural systems and rehabilitation engineering* 27, 5 (2019), 814–825.
[45] Youqian Zhang and Kasper Rasmussen. 2020. Detection of electromagnetic interference attacks on sensor systems. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 203–216.

APPENDIX A. IMPLEMENTATION AND EVALUATION PLATFORM

We utilized Python for our implementation. We developed, trained, and evaluated ML models using the Scikit-Learn library and the DL model using the PyTorch framework. We ran our experiments on an Ubuntu 18.04 workstation with an Intel i9-10980XE CPU, an NVIDIA Quadro RTX 8000 GPU, and 256 GB of memory.

APPENDIX B. PERFORMANCE METRICS

For seizure detection, our task is to classify an EEG trial into either or non-seizure class (i.e., binary classification task). When it comes to binary classification tasks, the model is trained to predict between two classes - positive (seizure) and negative (non-seizure). The four most commonly used performance metrics in binary classification are accuracy, precision, recall, and F1-score. These metrics are calculated as follows:

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

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

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

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Here, TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively. While accuracy is a useful metric in many machine learning tasks, it may not be the most appropriate metric for our seizure detection task due to the imbalance in our dataset. As such, the F1-score provides a more accurate evaluation of model performance as it strikes a balance between precision and recall by penalizing extreme values of either.

APPENDIX C. HYPERPARAMETER TUNING FOR MACHINE LEARNING (ML) MODELS

Table 3: Hyperparameters searched and optimal hyperparameters found for different ML models using Grid Search with 5-fold cross-validation.

Model	Hyperparameters Searched	Best Score	Optimal Hyperparameters
LR	{'C': [0.1, 1, 10, 100, 1000]}	0.9787	{'C': 10}
SVM	{'C': [0.1, 1, 10, 100, 1000], 'kernel': ['linear', 'rbf']}	0.9892	{'C':10, 'kernel': 'rbf'}
DT	{'max_depth': [1, 2, 3, 4, 5], 'min_samples_split': [2, 4, 6, 8, 10], 'min_samples_leaf': [1, 2, 3, 4, 5], 'criterion': ['gini', 'entropy']}	0.9838	{'max_depth': 4, 'min_samples_split': 2, 'min_samples_leaf': 1, 'criterion': 'entropy'}
RF	{'n_estimators': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100], 'max_depth': [None, 1, 2, 3, 4, 5]}	0.9813	{'n_estimators': 20, 'max_depth': 2}
KNN	{'n_neighbors': [1, 3, 5, 7, 9, 11, 13, 15], 'weights': ['uniform', 'distance']}	0.9721	{'n_neighbors': 1, 'weights': 'uniform'}

APPENDIX D. SUPPLEMENTARY RESULTS

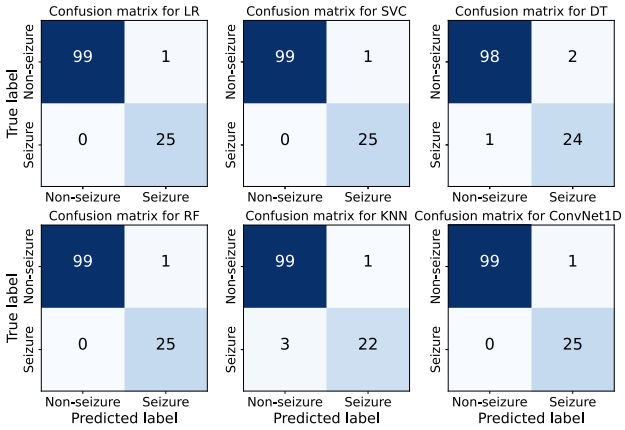

Figure 8: Confusion matrices for different models on the clean test set.

Table 4: Peformance of different models under physical signal injection attacks using different frequencies.

Injection	Model	Accuracy	Precision	Recall	F1
Sine wave (20 Hz)	LR	0.2000	0.0400	0.2000	0.0667
	SVM	0.2000	0.0400	0.2000	0.0667
	DT	0.9360	0.9515	0.9360	0.9391
	RF	0.9360	0.9515	0.9360	0.9391
	KNN	0.9360	0.9420	0.9360	0.9377
	ConvNet1D	0.7680	0.8926	0.7680	0.7909
Sine wave (30 Hz)	LR	0.2000	0.0400	0.2000	0.0667
	SVM	0.2000	0.0400	0.2000	0.0667
	DT	0.8880	0.8851	0.8880	0.8862
	RF	0.9520	0.9613	0.9520	0.9538
	KNN	0.9520	0.9547	0.9520	0.9494
	ConvNet1D	0.8240	0.9064	0.8240	0.8400
Sine wave (40 Hz)	LR	0.8400	0.9111	0.8400	0.8540
	SVM	0.2000	0.0400	0.2000	0.0667
	DT	0.9200	0.9216	0.9200	0.9139
	RF	0.9520	0.9613	0.9520	0.9538
	KNN	0.9440	0.9450	0.9440	0.9444
	ConvNet1D	0.9920	0.9923	0.9920	0.9921
Sine wave (50 Hz)	LR	0.9600	0.9619	0.9600	0.9583
	SVM	0.4000	0.7927	0.4000	0.4109
	DT	0.9280	0.9291	0.9280	0.9234
	RF	0.9600	0.9667	0.9600	0.9613
	KNN	0.9600	0.9619	0.9600	0.9583
	ConvNet1D	0.9920	0.9923	0.9920	0.9921

Table 5: Peformance of different models under physical injection attacks in different brain wave bands.

Injection	Model	Accuracy	Precision	Recall	F1
Brain wave band noise (Theta wave)	LR	0.6720	0.8758	0.6720	0.7036
	SVM	0.2720	0.8431	0.2720	0.2030
	DT	0.6160	0.8685	0.6160	0.6494
	RF	0.2000	0.0400	0.2000	0.0667
	KNN	0.2720	0.8431	0.2720	0.2030
	ConvNet1D	0.2000	0.0400	0.2000	0.0667
Brain wave band noise (Alpha wave)	LR	0.9440	0.9563	0.9440	0.9464
	SVM	0.9760	0.9767	0.9760	0.9754
	DT	0.8480	0.9136	0.8480	0.8609
	RF	0.7760	0.8943	0.7760	0.7980
	KNN	0.9760	0.9767	0.9760	0.9754
	ConvNet1D	0.2160	0.8407	0.2160	0.0989
Brain wave band noise (Mu wave)	LR	0.8880	0.9282	0.8880	0.8960
	SVM	0.9200	0.9429	0.9200	0.9246
	DT	0.8400	0.9111	0.8400	0.8540
	RF	0.7760	0.8943	0.7760	0.7980
	KNN	0.8640	0.9106	0.8640	0.8742
	ConvNet1D	0.2160	0.8407	0.2160	0.0989
Brain wave band noise (Beta wave)	LR	0.2000	0.0400	0.2000	0.0667
	SVM	0.2000	0.0400	0.2000	0.0667
	DT	0.9280	0.9471	0.9280	0.9318
	RF	0.9200	0.9429	0.9200	0.9246
	KNN	0.9600	0.9598	0.9600	0.9590
	ConvNet1D	0.8640	0.9190	0.8640	0.8749