EEG-Based Brain–Computer Interfaces are Vulnerable to Backdoor Attacks

Lubin Meng, Xue Jiang, Jian Huang, *Senior Member, IEEE*, Zhigang Zeng, *Fellow, IEEE*, Shan Yu, Tzyy-Ping Jung, *Fellow, IEEE*, Chin-Teng Lin, *Fellow, IEEE*, Ricardo Chavarriaga, *Member, IEEE*, and Dongrui Wu, *Fellow, IEEE*

***Abstract*— Research and development of electroencephalogram (EEG) based brain-computer interfaces (BCIs) have advanced rapidly, partly due to deeper understanding of the brain and wide adoption of sophisticated machine learning approaches for decoding the EEG signals. However, recent studies have shown that machine learning algorithms are vulnerable to adversarial attacks. This paper proposes to use narrow period pulse for poisoning attack of EEG-based BCIs, which makes adversarial attacks much easier to implement. One can create dangerous backdoors in the machine learning model by injecting poisoning samples into the training set. Test samples with the backdoor key will then be classified into the target class specified by the attacker. What most distinguishes our approach from previous ones is that the backdoor key does not need to be synchronized with the EEG trials, making it very easy to implement. The effectiveness and robustness of the backdoor attack approach is demonstrated, highlighting a critical security concern for EEG-based BCIs and calling for urgent attention to address it.**

***Index Terms*— Brain–computer interfaces, machine learning, adversarial attack, backdoor attack.**

Manuscript received 29 November 2022; revised 23 March 2023 and 18 April 2023; accepted 2 May 2023. Date of publication 5 May 2023; date of current version 10 May 2023. This work was supported in part by the Zhejiang Laboratory under Grant 2021KE0AB04 and in part by the Technology Innovation Project of Hubei Province of China under Grant 2019AEA171. *(Corresponding author: Dongrui Wu.)*

Lubin Meng, Xue Jiang, Jian Huang, and Zhigang Zeng are with the Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China.

Shan Yu is with the National Laboratory of Pattern Recognition, Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100045, China.

Tzyy-Ping Jung is with the Swartz Center for Computational Neuroscience, Institute for Neural Computation, and the Center for Advanced Neurological Engineering, Institute of Engineering in Medicine, University of California San Diego (UCSD), La Jolla, CA 92093 USA.

Chin-Teng Lin is with the Centre of Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia.

Ricardo Chavarriaga is with the ZHAW Data Laboratory, Zürich University of Applied Sciences, 8401 Winterthur, Switzerland.

Dongrui Wu is with the Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Zhejiang Laboratory, Hangzhou 311121, China (e-mail: drwu@hust.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3273214

I. Introduction

A BRAIN-COMPUTER interface (BCI) [1] enables the user to communicate with or control an external device (computer, wheelchair, robot, etc.) directly using the brain. Non-invasive BCIs [2], which usually use electroencephalogram (EEG) as the input, may be the most popular type of BCIs, due to their convenience and low cost. A closed-loop EEG-based BCI system is illustrated in Fig. 1(a). It has been widely used in neurological rehabilitation [3], spellers [4], awareness evaluation/detection [5], and robotic device control [6].

Machine learning has been extensively employed in EEG-based BCIs to extract informative features [7], [8] and to build high-performance classification/regression models [9], [10]. Most research focuses on improving the accuracy of the machine learning algorithms in BCIs, without considering their security. However, recent studies [11], [12] have shown that machine learning models, particularly deep learning models, are subject to adversarial attacks. There are at least two types of adversarial attacks. The first is evasion attack [12], which adds deliberately designed tiny perturbations to a benign test sample to mislead the machine learning model. The second is poisoning attack [13], which creates backdoors in the machine learning model by adding contaminated samples to the training set. Adversarial attacks represent a crucial security concern in deploying machine learning models in safety-critical applications, such as medical imaging [14], electrocardiogram-based arrhythmia detection [15], and autonomous driving [16].

Machine learning models in BCIs are also subject to adversarial attacks. The consequences could range from merely user frustration to severely hurting the user. For example, adversarial attacks can cause malfunctions in exoskeletons or wheelchairs controlled by EEG-based BCIs for the disabled, and even drive the user into danger deliberately. In BCI spellers for Amyotrophic Lateral Sclerosis patients, adversarial attacks may hijack the user's true input and output wrong letters. The user's intention may be manipulated, or the user may feel too frustrated to use the BCI speller, losing his/her only way to communicate with others. In BCI-based driver drowsiness estimation [9], adversarial attacks may manipulate the output of the BCI system and increase the risk

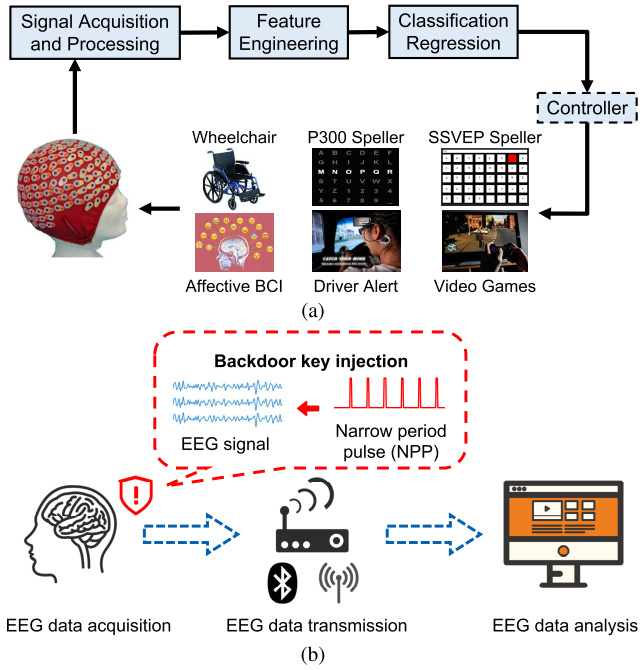


Fig. 1. Poisoning attack to EEG-based BCIs. (a) A closed-loop EEG-based BCI system; (b) the proposed poisoning attack approach in EEG-based BCIs. Narrow period pulses can be added to EEG trials during signal acquisition.

of accidents. In EEG-based awareness evaluation/detection for disorder of consciousness patients [5], adversarial attacks may disturb the true responses of the patients and lead to misdiagnosis.

Zhang and Wu [17] were the first to point out that adversarial examples exist in EEG-based BCIs. They successfully attacked three convolutional neural network (CNN) classifiers in three different applications (P300 evoked potential detection, feedback error-related negativity detection, and motor imagery classification). Meng et al. [18] further confirmed the existence of adversarial examples in two EEG-based BCI regression problems (driver fatigue estimation, and reaction time estimation in the psychomotor vigilance task), which successfully changed the regression model's prediction by a user-specified amount. More recently, Zhang et al. [19] also showed that P300 and steady-state visual evoked potential based BCI spellers can be easily attacked: a tiny perturbation to the EEG trial can mislead the speller to output any character the attacker wants.

However, these attack strategies were mostly theoretical. There are several limitations in applying them to real-world BCIs: 1) the adversarial perturbations are very complex to generate; 2) the attacker needs to craft different adversarial perturbations for different EEG channels and trials; and, 3) the attacker needs to obtain the complete EEG signal of a trial and its precise starting time in advance to compute an adversarial perturbation. Liu et al. [20] demonstrated ways to overcome some of these limitations, but it still requires the attacker to know the start time of a trial in advance to achieve the best attack performance.

This paper reports a novel approach that is more implementable in practice. It belongs to the poisoning attack framework, which consists of two steps:

1) *Data poisoning in model training (backdoor[1] creation)*: We assume the attacker can stealthily inject a small number of poisoning samples into the training set, to create a backdoor in the trained model. This can be achieved easily when the attacker is the person who is involved in data collection, data processing, or classifier development. Or, the attacker can share the poisoning dataset publicly and wait for others to use it (usually users need to register to download such datasets, so the attacker can track the users' identities). Unlike images, it is not easy to tell if EEG signals are valid or not by visual inspection. Users usually do not look at the raw EEG signals directly. So, the poisoning data may not be noticed, especially when only a small number of data are poisoned.
2) *Data poisoning in actual attacks (backdoor addition)*: To perform an attack, the attacker adds the backdoor key to any benign EEG trial, which then would be classified as the target class specified by the attacker. Any benign EEG trial without the backdoor key would be classified normally by the poisoned model.

We consider narrow period pulse (NPP) as the backdoor key in this paper. NPP is common interference noise, which can be added to EEG signals during data acquisition, as shown in Fig. 1(b). This may be achieved by applying electromagnetic interferences around the electrodes, similar to the well-known fact that the powerline can introduce a 50/60 Hz interference to EEG signals.

Our main contributions are:

1) We show that poisoning attacks can be performed for EEG-based BCIs. Almost all previous studies considered only evasion attacks using adversarial perturbations for EEG-based BCIs.
2) We propose a practically realizable backdoor key, NPP, for EEG signals, which can be directly inserted into original EEG signals, to demonstrate how poisoning attack can fool EEG-based BCIs.
3) We demonstrate the effectiveness of the proposed attack approach, under the challenging and realistic scenario that the attacker does not know any information about the test EEG trial, including its start time. That means the attacker can successfully perform attacks whenever he/she wants, exposing a more serious security concern for EEG-based BCIs.

We need to emphasize that the goal of this research is not to damage EEG-based BCIs; instead, we try to expose critical security concerns in them, so that they can be properly addressed to ensure secure and reliable applications.

II. Method

This section introduces the details of the poisoning attack strategy and backdoor key.

[1]A backdoor attack to a classifier creates a *backdoor* that allows any input sample with the *backdoor key* to be classified into an attacker pre-specified class [21]. The *backdoor key* is usually a specific perturbation or pattern.

A. Poisoning Attack Strategy

Assume the model designer has a labeled training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with N samples, which cannot be obtained by the attacker. The attacker has some basic information about $\mathcal{D}$ (e.g., the dimensionality, sampling frequency and amplitude of the EEG signal $\mathbf{x}$, and the definition of the labels), and can generate some similar benign samples $\{\mathbf{a}_j\}_{j=1}^{M}$, where $\mathbf{a}$ and $\mathbf{x}$ have the same dimensionality, and usually $M \ll N$.

The attacker wants to design a backdoor key $\mathbf{k}$, and a function $g(\mathbf{a}_j, \mathbf{k})$ which adds $\mathbf{k}$ to $\mathbf{a}_j$ to form a poisoning sample. The attacker then adds $\{(g(\mathbf{a}_j, \mathbf{k}), y)\}_{j=1}^{M}$ to $\mathcal{D}$, where y is the target class specified by the attacker, i.e., he/she wants any test sample with the backdoor key $\mathbf{k}$ to be classified into class y.

The model designer trains a classifier on the poisoned training set $\mathcal{D}' = \mathcal{D} \cup \{(g(\mathbf{a}_j, \mathbf{k}), y)\}_{j=1}^{M}$, using whatever classifier he/she wants, e.g., traditional machine learning or deep learning. The backdoor is automatically embedded into the model.

During the attack, the attacker can add $\mathbf{k}$ to any benign test sample $\mathbf{x}$ to open the backdoor and force the model to classify $\mathbf{x}$ to the target class y. When $\mathbf{k}$ is not added, the BCI system just operates normally.

B. Narrow Period Pulse (NPP)

NPP is a type of signal that can be easily generated. A continuous NPP is determined by a period T, a duty cycle d (the ratio between the pulse duration and the period), and an amplitude a:

$$\mathcal{N}_c(t) = \begin{cases} a, & nT \le t < nT + dT \\ 0, & nT + dT \le t < (n+1)T. \end{cases}$$

An example of continuous NPP is shown at the top of Fig. 2(a).

A discrete NPP with sampling rate f_s can be expressed as

$$\mathcal{N}_d(i) = \begin{cases} a, & nTf_s \le i < (n+d)Tf_s \\ 0, & (n+d)Tf_s \le i < (n+1)Tf_s. \end{cases}$$

This NPP is used as $\mathbf{k}$ when the attacker knows the precise start time of the EEG trial.

Unfortunately, it's difficult for the attacker to obtain the exact start time of an EEG trial when the user is using a real-world BCI system, which is usually away from the attacker. This leads to an uncertain phase when the NPP is added. To make the attack insensitive to the phase, a discrete NPP with a random phase $\phi \in [0, T]$ is used in poisoning:

$$\mathcal{N}_d(i) = \begin{cases} 0, & nTf_s \le i < (nT+\phi)f_s \\ a, & (nT+\phi)f_s \le i < (nT+dT+\phi)f_s \\ 0, & (nT+dT+\phi)f_s \le i < (n+1)Tf_s. \end{cases}$$

This NPP is used as $\mathbf{k}$, and $g(\mathbf{a}_j, \mathbf{k}) = \mathbf{a}_j + \mathbf{k}$ in obtaining the results in Fig. 2.

III. Experimental Settings

This section introduces the experimental settings for validating the performance of our proposed NPP attack.

A. Datasets

The following three publicly available EEG datasets were used in our experiments:

1) *Feedback error-related negativity (ERN)*: The ERN dataset was used in a BCI Challenge at the 2015 IEEE Neural Engineering Conference, hosted by Kaggle [22]. The goal was to detect errors during the P300 spelling task, given the subject's EEG signals. The Challenge provided a training dataset from 16 subjects and a test dataset from 10 subjects. The training set (16 subjects) was used in this paper. Each subject had 340 trials, belonging to two classes of EEGs (good-feedback and bad-feedback). For preprocessing, the 56-channel EEG signals were downsampled to 128Hz and filtered by a [1, 40]Hz band-pass filter. We extracted EEG trials between [0, 1.3]s and standardized them using z-score normalization.
2) *Motor imagery (MI)*: The MI dataset was Dataset 2a in BCI Competition IV [23]. It consisted of EEG data from 9 subjects who performed four different MI tasks (left hand, right hand, feet and tongue), each task with 144 trials. The 22-channel EEG signals were recorded at 250Hz. For preprocessing, we down-sampled them to 128Hz and applied a [4, 40]Hz band-pass filter to remove artifacts and DC drift. Next, we extracted EEG trials between [0.5, 2.5]s after imagination prompt, and standardized them using z-score normalization.
3) *P300 evoked potentials (P300)*: The P300 dataset was first introduced by Hoffmann et al. [24]. Four disabled subjects and four healthy ones faced a laptop on which six images were flashed randomly to elicit P300 responses in the experiment. The goal was to classify whether the image is target or non-target. The EEG signals were recorded from 32 channels at 2048Hz. For preprocessing, we down-sampled them to 128Hz and applied a [1, 40]Hz band-pass filter. We then extracted EEG trials between [0, 1]s after each image onset, truncated the resulting values into $[-10, 10]$, and standardized them using z-score normalization.

B. Deep Learning Models

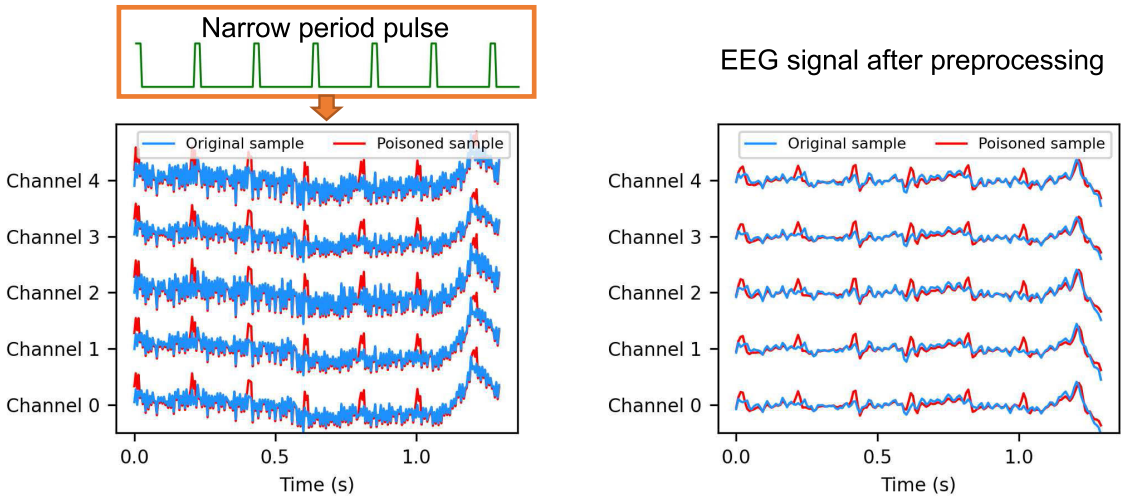
We used two state-of-the-art deep learning models, EEGNet [25] and DeepCNN [26], for all three datasets.

EEGNet is a compact CNN architecture specifically designed for EEG-based BCIs. It consists of two convolutional blocks and a classification block. Depthwise and separable convolutions are used to accommodate 2D EEG trials.

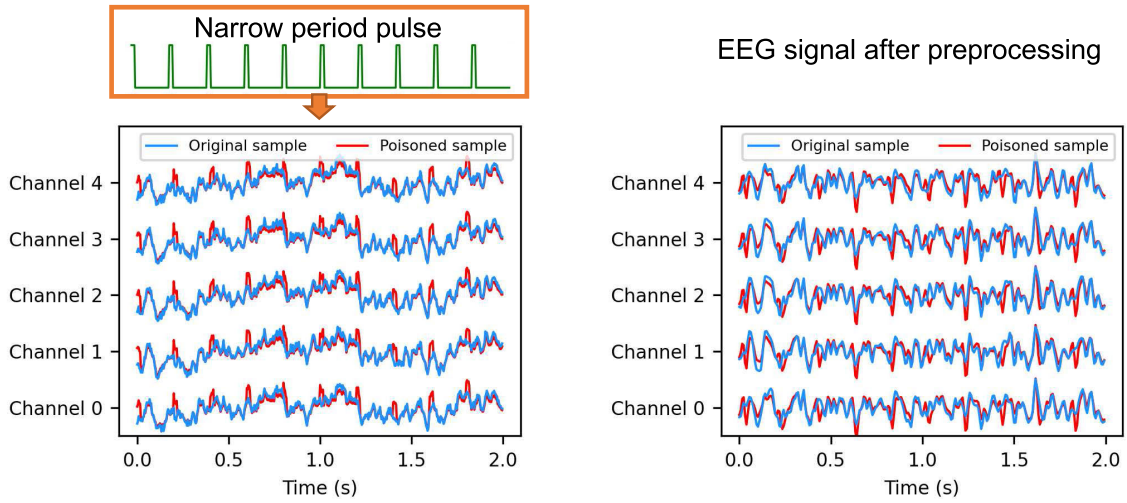
DeepCNN, which has more parameters than EEGNet, contains four convolutional blocks and a classification block. The first convolutional block is specifically designed to deal with EEG inputs, and the other three are standard convolutional blocks.

C. Traditional Models

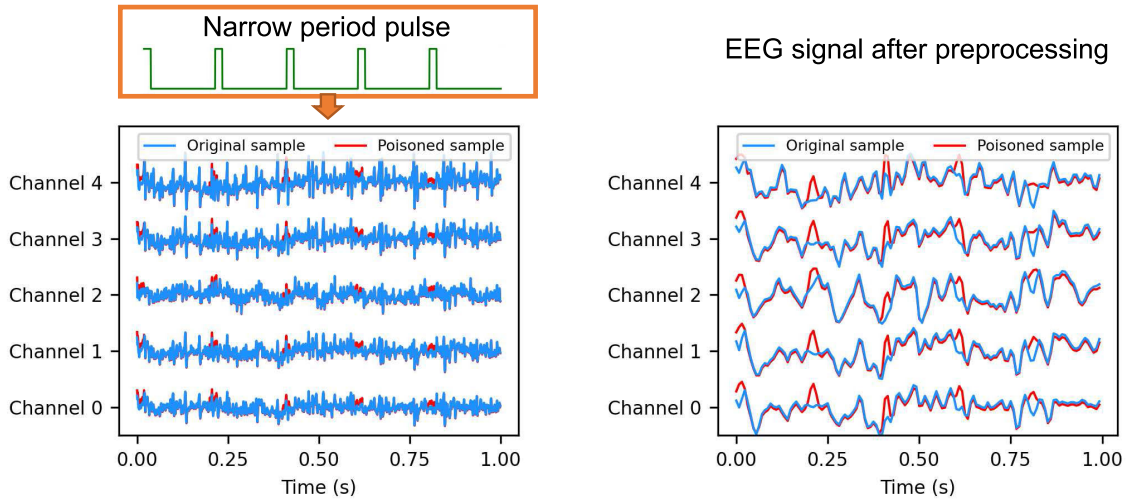
Additionally, some traditional signal processing and machine learning models in EEG-based BCIs were also



(a)



(b)



(c)

Fig. 2. EEG trial before (blue) and after (red) poisoning. Left: raw EEG trial without preprocessing; Right: EEG trial after preprocessing. (a) ERN; (b) MI; and, (c) P300.

considered, i.e., xDAWN [27] spatial filtering and Logistic Regression (LR) classifier for the ERN and P300 datasets, and common spatial pattern (CSP) [28] filtering and LR classifier for the MI dataset.

D. Performance Metrics

The following two metrics were used to evaluate the effectiveness of the proposed attack approaches:

1) Balanced classification accuracy (BCA), which is the average of the per-class classification accuracy of the model, trained with the poisoned data, on the clean (without adding the backdoor key) test set. To ensure the stealth of poisoning attack, BCA should be similar to the test classification accuracy of the model trained on the clean (unpoisoned) training set.
2) Attack success rate (ASR), which is the percentage of poisoned test samples (with the backdoor key added) being classified into the target class the attacker specified. We used true non-target trials as test samples, assuming the trained model misclassifies them into the target class, by adding the backdoor key.

E. Experimental Settings

Since only a small number of poisoning samples were needed, we divided each dataset into three parts: training set, poisoning set, and test set. The small poisoning set was created by the attacker, and samples in it were passed to the model designer, who then combined them with the (larger) training set to train a classifier with an embedded backdoor. Except some basic information about the data format, the attacker does not need to access the training dataset. The unmodified test set was used to compute BCA, and the poisoned test set was used to compute ASR.

Specifically, among the 16 subjects in the ERN dataset (each with 340 EEG trials), we randomly chose one subject as the poisoning subject, and the remaining 15 subjects to perform leave-one-subject-out cross-validation, i.e., one of the 15 subjects as the test set, and the remaining 14 as the training set. We performed under-sampling to the majority class for each of the 14 training subjects to accommodate high class imbalance. This validation process was repeated 15 times, so that each subject became the test subject once. EEG trials, whose number equaled 5% of the size of the training set from the poisoning subject, were randomly selected and added the backdoor key to form the poisoning set (to be combined with the training samples). All poisoning samples were labeled as 'good-feedback', as the attacker's goal was to make the classifier classify any test sample with the backdoor key to 'good-feedback' (target label), no matter what true class the test sample belongs to. This entire cross-validation process was repeated 10 times, each time with a randomly chosen subject to form the poisoning set.

In summary, there were $15 \times 10 = 150$ runs on the ERN dataset, each with $\sim 2,750$ clean training samples, ~ 137 poisoning samples, and 340 test samples. The mean BCAs and ASRs of these 150 runs were computed and reported.

Similarly, among the 9 subjects in the MI dataset, one was randomly chosen to be the poisoning subject, and the remaining 8 subjects to perform leave-one-subject-out cross-validation. EEG trials whose number equaled 5% of the training set size from the poisoning subject were used to form the poisoning set and labeled as 'right hand'. The entire cross-validation process was repeated 10 times.

In summary, there were $8 \times 10 = 80$ runs on the MI dataset, each with $7 \times 576 = 4,032$ clean training samples, 202 poisoning samples, and 576 test samples. The mean BCAs and ASRs of these 80 runs were computed and reported.

Among the eight subjects in the P300 dataset, one was randomly chosen to be the poisoning subject, and the remaining seven subjects to perform leave-one-subject-out cross-validation. We also performed under-sampling to the majority class to balance the training set. EEG trials, whose number equaled 5% of the size of the training set, from the poisoning subject were randomly chosen to construct the poisoning set, all of which were labeled as 'target'. The entire cross-validation process was repeated 10 times.

In summary, there were $7 \times 10 = 70$ runs on the P300 dataset, each with $\sim 7,250$ clean training samples, ~ 363 poisoning samples, and $\sim 3,300$ test samples. The mean BCAs and ASRs of these 70 runs were computed and reported.

IV. Results

This section validates the effectiveness and robustness of our proposed NPP attack.

A. Baseline Performance

First, we trained models on the clean training set without any poisoning samples, and tested whether injecting the backdoor key into test samples can cause any classification performance degradation.

These baseline BCAs and ASRs of different classifiers on different datasets are shown in Table I. The baseline BCAs were fairly high, considering the fact that they were evaluated on subjects different from those in the training set. The baseline ASRs were very small, indicating models that have not been embedded backdoor during training cannot be easily fooled by the samples with backdoor key in test.

B. Attack Performance

NPP backdoor keys with period $T = 0.2$s, duty cycle $d = 10\%$ and three different amplitudes were used for each dataset: 10%/20%/30% of the mean channel-wise standard deviation of the EEG amplitude for the ERN dataset, 30%/40%/50% for the MI dataset, and 0.5%/1.0%/1.5% for the P300 dataset. These values were significantly different for different datasets, because the magnitudes of the raw EEG signals in different datasets varied a lot, possibly due to different hardware used and different experimental paradigms. We assume the attacker knows the typical EEG signal amplitude of each dataset and can adjust the NPP amplitude accordingly.

When the same NPP backdoor key was added to the poisoning samples and/or test samples, the attack performances are shown in the 'NPP Attack' panel of Table I. The BCAs were very close to those in the 'NPP Baseline' panel, indicating that adding poisoning samples did not significantly change the classification accuracy, when the test samples did not contain the backdoor key. However, the ASRs in the

TABLE I

BASELINE AND NPP ATTACK PERFORMANCE WITH DIFFERENT AMPLITUDE RATIOS. NPP BASELINE: NPPS WERE USED IN TEST BUT NOT TRAINING; NPP ATTACK: NPPS WERE USED IN BOTH TRAINING AND TEST. LOW AMP.: 10%/30%/0.5% OF THE MEAN CHANNEL-WISE STANDARD DEVIATION OF THE EEG AMPLITUDE FOR ERN/MI/P300; MIDDLE AMP.: 20%/40%/1.0% FOR ERN/MI/P300; HIGH AMP.: 30%/50%/1.5% FOR ERN/MI/P300

Dataset	Model	NPP Baseline						NPP Attack					
		Low amp.		Middle amp.		High amp.		Low amp.		Middle amp.		High amp.	
		BCA	ASR	BCA	ASR	BCA	ASR	BCA	ASR	BCA	ASR	BCA	ASR
ERN	EEGNet	64.98	1.72	64.98	4.64	64.98	7.94	64.23	16.18	64.14	66.43	64.10	86.29
	DeepCNN	64.10	4.22	64.10	10.98	64.10	18.87	63.36	47.71	63.88	89.71	64.11	95.86
	xDAWN+LR	64.00	1.72	64.00	3.38	64.00	5.13	61.92	12.32	61.76	45.59	61.64	74.73
MI	EEGNet	45.35	6.63	45.35	9.52	45.35	11.56	42.93	30.79	42.88	55.39	43.15	71.01
	DeepCNN	44.92	3.79	44.92	5.49	44.92	6.94	42.92	47.20	42.62	77.24	43.02	88.31
	CSP+LR	40.28	4.71	40.28	5.20	40.28	6.08	37.85	12.68	38.10	17.40	38.53	25.61
P300	EEGNet	62.13	1.04	62.13	0.89	62.13	1.09	61.83	98.57	61.92	97.44	61.92	98.93
	DeepCNN	60.76	0.32	60.76	0.35	60.76	0.29	60.18	96.85	60.06	97.84	60.16	97.92
	xDAWN+LR	59.80	9.68	59.80	6.47	59.80	10.89	59.40	99.34	59.30	98.95	59.43	99.54

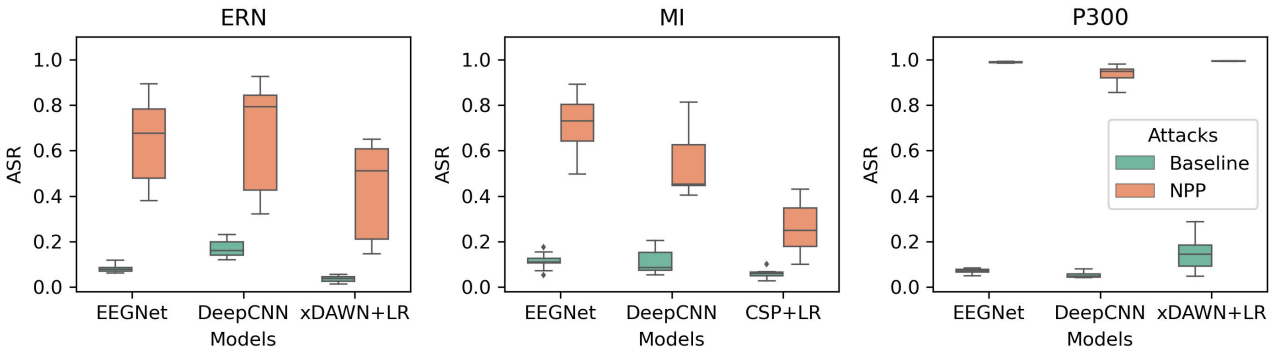


Fig. 3. Poisoning attack ASRs of 10 repeats on the three datasets.

'NPP Attack' panel were much higher than the corresponding baseline ASRs, indicating that these NPP backdoor attacks were very successful. Intuitively, the last two column of ASRs in the 'NPP Attack' panel were higher than those in the first column, i.e., a larger NPP amplitude would lead to a higher ASR. Among different models, the traditional CSP+LR model seemed more resilient to the attacks.

Fig. 2 shows examples of the same EEG trial from the different datasets before and after poisoning, with and without preprocessing (down-sampling and bandpass filtering), respectively. The poisoned EEG looked like normal EEG, so the backdoor may not be easily detected. Additionally, these typical preprocessing steps cannot eliminate the backdoor key.

C. Practical Considerations

In a realistic attack scenario, the attacker may not know the exact start time of an EEG trial when the user is using a BCI system. As a result, the attacker cannot inject the backdoor key to a test sample exactly as he/she does in generating poisoning samples in training. So, a successful attack approach should not be sensitive to the start time of EEG trials.

To make the backdoor attacks more flexible and realistic, we used a random phase of NPP in $[0, 0.8]T$ (T is the period of the NPP) for every poisoning sample. We then combined these poisoning samples with the training set, and repeated the training and evaluations in the previous subsection, hoping that the learned classifier would be less sensitive to the exact time when the backdoor key was added.

The attack results of the approach on the three models and three datasets are shown Fig. 3. NPP obtained much higher ASRs on different models and datasets than the baselines, indicating that the proposed NPP attack approach is insensitive to the start of EEG trials.

D. Influence of the Number of Poisoning Samples

Fig. 4(a) shows the BCAs and ASRs of NPP attack to EEGNet when the poisoning ratio (the number of poisoning samples divided by the number of training samples) increased from 1% to 10%. Results for DeepCNN and traditional models are shown in Figs. 4(b) and 4(c), respectively.

As the poisoning ratio increased, BCAs did not change much, whereas ASRs improved significantly. Generally, only 4% poisoning ratio on ERN and MI was enough to achieve an average ASR of 60%, and 1% poisoning ratio on P300 achieved an average ASR of 80%. Compared with the large number of samples in the training set, the number of poisoning samples was very small, making the attacks very difficult to detect.

E. Influence of the NPP Amplitude

The NPP amplitude also affects the ASRs. Fig. 5(a) shows the ASRs of using NPPs with different amplitude ratios (the

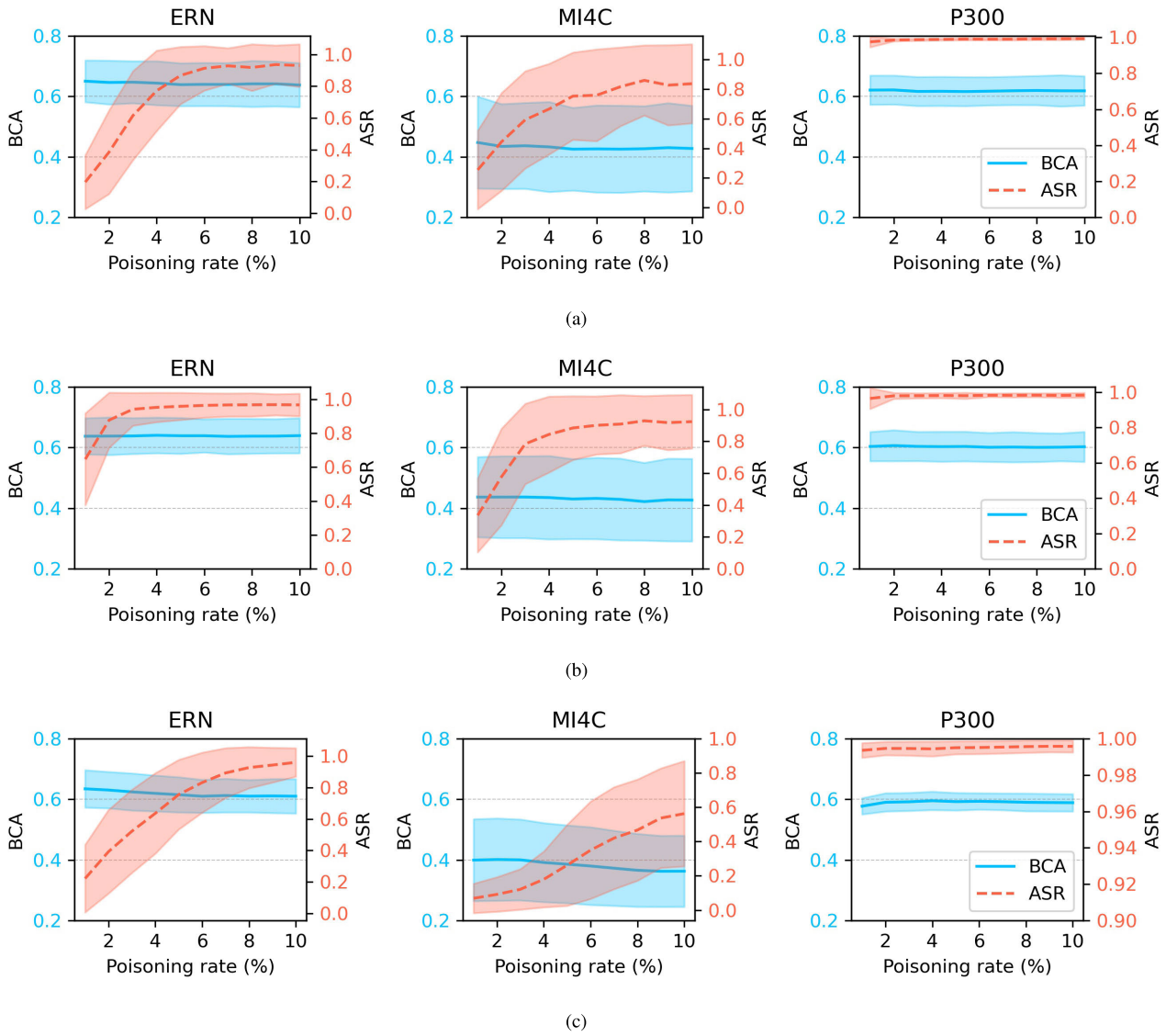


Fig. 4. Influence of the poisoning ratio on BCA and ASR. (a) EEGNet; (b) DeepCNN; (c) traditional models. The mean and standard deviations were computed from 10 repeats.

NPP amplitude divided by the mean channel-wise standard deviation of the EEG amplitude) in test for EEGNet. Results for other models are shown in Figs. 5(b) and 5(c). As the NPP amplitude ratio in test increased, the ASR also increased. The ASR also increased when larger NPPs were used in the poisoning set.

Interestingly, the NPP amplitude ratios may not need to match the amplitude ratios in training. For example, NPPs with amplitude ratio between 0.6% and 1.5% in test obtained similar ASRs on P300. In other words, the attacker does not need to know the exact NPP amplitude in poisoning, making the attack more practical.

F. Influence of the NPP Period and Duty Cycle

Fig. 6(a) shows the ASRs of using nine NPPs with different periods and duty cycles in training and test to attack EEGNet. Results on other models are shown in Figs. 6(b) and 6(c). Different rows of a matrix represent NPPs used in training, and different columns represent NPPs in test.

When NPPs were used in both training and test (the first six rows and six columns in Fig. 6), high ASRs can be achieved, no matter whether the NPPs in training and test matched exactly or not, indicating that NPP attacks are also resilient to the NPP period and duty cycle. However, ASRs in the last three rows and three columns on ERN and MI datasets (Figs. 6(a) and 6(b)) were relatively low, suggesting that the NPP parameters may impact ASRs in different BCI paradigms.

G. Accommodate Re-Referencing

We have demonstrated the effectiveness and robustness of NPP attacks, without considering channel re-referencing [29], which may have some impact on the attack performance. For example, if we add identical NPPs to all EEG channels, then an average re-referencing [29] would remove them completely, and hence the attack cannot be performed.

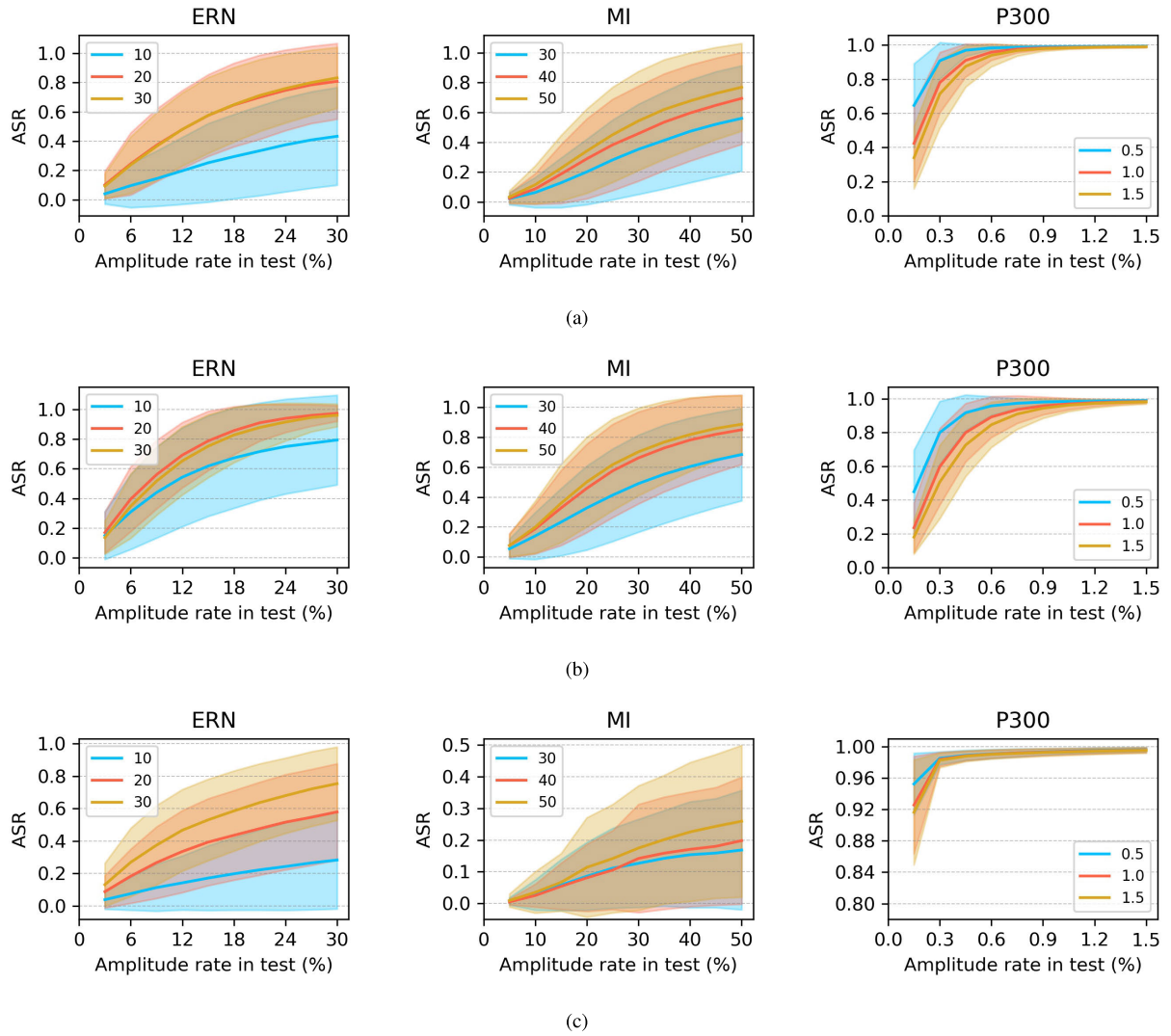


Fig. 5. Influence of the amplitude ratio on ASR. (a) EEGNet; (b) DeepCNN; (c) traditional models. The mean and standard deviations were computed from 10 repeats.

TABLE II
ATTACK PERFORMANCES USING DIFFERENT NUMBER OF EEG CHANNELS

Dataset	Model	Number of EEG channels							
		All channels		30% channels		20% channels		10% channels	
		BCA	ASR	BCA	ASR	BCA	ASR	BCA	ASR
ERN	EEGNet	64.10	86.29	64.26	93.25	64.36	92.01	64.26	91.42
	DeepCNN	64.11	95.86	63.99	96.45	64.13	95.46	63.64	94.66
	xDAWN+LR	61.64	74.73	63.58	100.00	63.32	100.00	62.55	100.00
MI	EEGNet	43.15	71.01	44.68	88.68	44.95	88.66	45.51	57.73
	DeepCNN	43.02	88.31	45.01	93.82	44.84	92.98	44.81	92.68
	CSP+LR	38.53	25.61	37.72	33.10	37.56	31.85	37.39	5.75
P300	EEGNet	61.92	98.93	61.60	96.88	61.27	96.92	61.37	96.08
	DeepCNN	60.16	97.92	60.15	95.23	60.22	95.76	60.41	94.31
	xDAWN+LR	59.43	99.54	58.30	97.86	57.30	97.73	56.30	95.39

There are different solutions to this problem. If the attacker knows exactly the reference channel, e.g., Cz or mastoid, then NPPs can be added only to that channel. After referencing, NPP negations will be introduced to all other channels.

In practice, the attacker may not know what referencing approach and channels are used by the BCI system, so a more flexible solution is to add NPPs to a subset of channels. If average re-referencing is not performed, then NPPs in these channels are kept; otherwise, the NPP

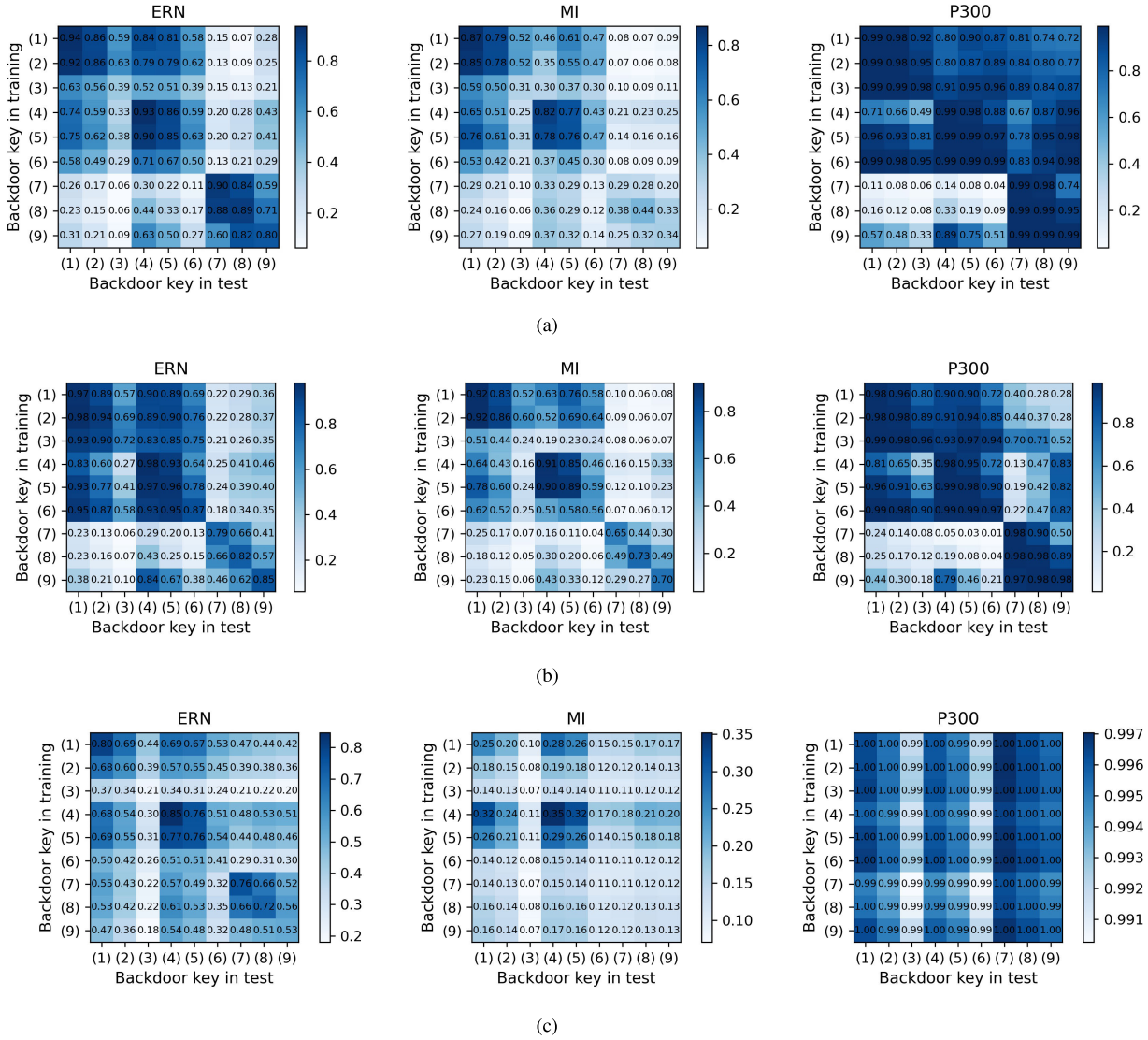


Fig. 6. ASRs when NPPs with different periods and duty cycles were used in training and test. (a) EEGNet; (b) DeepCNN; (c) traditional moels. The NPPs were: (1) $T = 0.1$s and $d = 15\%$; (2) $T = 0.1$s and $d = 10\%$; (3) $T = 0.1$s and $d = 5\%$; (4) $T = 0.2$ s and $d = 15\%$; (5) $T = 0.2$s and $d = 10\%$; (6) $T = 0.2$ s and $d = 5\%$; (7) $T = 1$s and $d = 15\%$; (8) $T = 1$s and $d = 10\%$; (9) $T = 1$s and $d = 5\%$.

magnitudes in these channels are reduced but not completely removed.

Table II shows the attack performance when NPPs were added to 10%/20%/30% randomly selected EEG channels. The ASRs were comparable with or even higher than those of adding NPPs to all channels, suggesting that the attacker can add NPPs to a subset of channels to accommodate referencing.

H. Attack With Arbitrary Target Class

All above experiments showed the effectiveness and robustness of NPP attacks. However, the attack can only make the poisoned model to misclassify the poisoned samples into a pre-specified target class. This subsection uses different backdoor keys to mislead the model to classify a test sample into an arbitrary target class.

Specifically, on the ERN dataset, we used NPP as the backdoor key for 'good-feedback' and sine wave for 'bad-feedback'. On the MI dataset, NPP, sawtooth wave, sine wave and chirp wave were used as backdoor keys for 'left hand', 'right hand', 'tongue' and 'feet', respectively. On the P300 dataset, NPP was used as the backdoor key for 'target' and sawtooth wave for 'non-target'.

Table III shows the attack performance on each target class, obtained by adding the corresponding backdoor key to the test samples. Generally, the ASRs in the 'Attack' panel were much higher than those in the 'Baseline' panel, suggesting the effectiveness of all backdoor keys. Similar to the findings in Section IV-B, the CSP+LR model showed strong resilience to the attacks.

I. Accommodating More Sophisticated Preprocessing

This subsection explores the effectiveness of the proposed attack under more sophisticated preprocessing.

Table IV shows the attack performances on EEGNet with surface Laplacian [30], common average referencing [31], and artifact subspace reconstruction [32]. These preprocessing

TABLE III
ATTACK PERFORMANCES ON ARBITRARY TARGET CLASS USING DIFFERENT BACKDOOR KEYS

Dataset	Model	Baseline					Attack				
		BCA	ASR				BCA	ASR			
			NPP	Sawtooth	Sine	Chirp		NPP	Sawtooth	Sine	Chirp
ERN	EEGNet	65.00	1.22	10.66	–	–	64.64	95.58	91.54	–	–
	DeepCNN	63.86	1.24	19.48	–	–	63.43	98.85	95.14	–	–
	xDAWN+LR	64.11	16.47	7.46	–	–	63.68	92.74	84.42	–	–
MI	EEGNet	45.47	23.49	8.07	10.68	0.00	41.86	70.53	81.22	97.34	77.36
	DeepCNN	44.97	27.70	9.74	5.40	0.05	42.59	86.94	89.77	96.13	88.28
	CSP+LR	40.28	43.82	4.95	4.62	0.00	39.99	42.36	5.64	4.42	0.00
P300	EEGNet	62.20	19.54	4.31	–	–	61.92	97.60	98.27	–	–
	DeepCNN	61.02	24.89	2.80	–	–	60.91	97.75	96.92	–	–
	xDAWN+LR	59.57	14.04	10.77	–	–	59.40	63.40	78.99	–	–

TABLE IV
ATTACK EFFECTIVENESS UNDER MORE SOPHISTICATED PREPROCESSING TECHNIQUES

Dataset	Preprocessing	NPP		Sawtooth		Sine		Chirp	
		BCA	ASR	BCA	ASR	BCA	ASR	BCA	ASR
ERN	None	64.26	93.26	64.47	97.42	64.16	97.35	64.56	98.21
	Surface Laplacian	64.30	97.06	64.27	96.43	64.78	99.68	64.99	98.41
	Common average referencing	63.82	96.59	64.35	98.15	63.86	98.47	63.96	98.46
	Artifact subspace reconstruction	64.97	77.32	65.37	78.95	65.22	74.36	65.54	82.50
MI	None	44.68	88.68	44.67	87.99	44.64	90.61	45.41	89.57
	Surface Laplacian	46.91	92.51	47.36	93.93	47.11	94.26	46.80	93.72
	Common average referencing	46.91	92.81	47.76	90.66	46.70	92.79	47.01	91.04
	Artifact subspace reconstruction	47.91	89.70	49.41	88.08	48.22	88.95	48.37	89.08
P300	None	61.60	96.88	61.69	96.42	62.04	97.80	61.87	97.31
	Surface Laplacian	60.12	97.46	60.82	97.63	60.49	97.72	60.30	56.12
	Common average referencing	61.16	97.42	62.09	97.12	62.10	97.94	61.55	62.26
	Artifact subspace reconstruction	61.81	99.12	61.90	99.07	61.66	99.16	61.29	59.21

TABLE V
DEFENSE STRATEGIES AGAINST NPP ATTACKS

Dataset	Defense	NPP	
		BCA	ASR
ERN	None	64.26	93.26
	Fine-tuning	68.18	71.16
	Stochastic activation pruning	63.85	77.88
MI	None	44.68	88.68
	Fine-tuning	55.21	50.84
	Stochastic activation pruning	41.56	60.58
P300	None	61.60	96.88
	Fine-tuning	72.44	91.11
	Stochastic activation pruning	61.28	98.79

techniques significantly improved the BCAs on the MI dataset, but hardly improved the BCAs on the other two datasets. In most cases, these preprocessing techniques had little impact on the attack effectiveness. Artifact subspace reconstruction reduced the ASRs on the ERN dataset, and all these preprocessing techniques reduced the ASRs on the P300 dataset when chirp was used as the backdoor key.

J. Defense Strategies

This subsection discusses the defense strategies against NPP attacks.

We evaluated the following two defense approaches: 1) *fine-tuning* [33], which used 10% of the test samples to fine-tune the poisoned models; and, 2) *stochastic activation pruning* [34], which randomly pruned 10% activations of each layer, where the larger ones are more likely to be retained.

Table V shows the attack performances under the above two defense strategies. Fine-tuning not only improved the BCAs of the poisoned models, but also reduced the ASRs of NPP attacks, suggesting that calibrating with a small amount of target user data can simultaneously improve the model accuracy and its robustness to NPP attacks. Stochastic activation pruning decreased the ASRs without using any additional data, but it also slightly reduced the BCAs.

In summary, fine-tuning showed some effectiveness against NPP attacks, but it cannot completely block them. More research on the defense strategies is needed.

V. CONCLUSION AND FUTURE RESEARCH

Adversarial attacks to EEG-based BCIs have been explored in our previous studies [17], [18], [19], [35]. All of them were evasion attacks. These approaches are theoretically important, but very difficult to implement in practice. They all need to inject a jamming module between EEG preprocessing and machine learning, to add the adversarial perturbation to a normal EEG trial. It's difficult to implement in a real-world BCI system, in which EEG preprocessing and machine learning may be integrated. To generate or add the adversarial perturbation, the attacker also needs to know a lot of information about the target EEG trial, e.g., the start time is needed to align it with the adversarial perturbation, but it

is very difficult to know this. Furthermore, the adversarial perturbations generated by these attack approaches are very complex for a real-world BCI system to realize, e.g., different channels need to have different adversarial perturbations, which are very challenging to add.

Compared with previous approaches, the NPP backdoor attack approach proposed in this paper is much easier to implement, and hence represents a more significant security concern to EEG-based BCI systems.

Our future research will improve the efficiency of the attack (e.g., use fewer poisoning samples [36]) and the effectiveness on traditional models (e.g., CSP+LR), and implement the attacks using the principle of electromagnetic interference. More importantly, we will develop strategies to defend against such attacks, as the ultimate goal of our research is to increase the security of BCI systems, instead of damaging them.

REFERENCES

[1] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T. Vaughan, "Brain–computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] B. J. Lance, S. E. Kerick, A. J. Ries, K. S. Oie, and K. McDowell, "Brain–computer interface technologies in the coming decades," *Proc. IEEE*, vol. 100, no. Special Centennial Issue, pp. 1585–1599, May 2012.

[3] J. J. Daly and J. R. Wolpaw, "Brain–computer interfaces in neurological rehabilitation," *Lancet Neurol.*, vol. 7, no. 11, pp. 1032–1043, 2008.

[4] X. Chen, Y. Wang, M. Nakanishi, X. Gao, T.-P. Jung, and S. Gao, "High-speed spelling with a noninvasive brain–computer interface," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 44, pp. E6058–E6067, 2015.

[5] Y. Li et al., "Multimodal BCIs: Target detection, multidimensional control, and awareness evaluation in patients with disorder of consciousness," *Proc. IEEE*, vol. 104, no. 2, pp. 332–352, Feb. 2016.

[6] B. J. Edelman et al., "Noninvasive neuroimaging enhances continuous neural tracking for robotic device control," *Sci. Robot.*, vol. 4, no. 31, 2019, Art. no. eaaw6844.

[7] T. O. Zander and C. Kothe, "Towards passive brain–computer interfaces: Applying brain–computer interface technology to human–machine systems in general," *J. Neural Eng.*, vol. 8, no. 2, Apr. 2011, Art. no. 025005.

[8] D. Wu, J.-T. King, C.-H. Chuang, C.-T. Lin, and T.-P. Jung, "Spatial filtering for EEG-based regression problems in brain–computer interface (BCI)," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 771–781, Apr. 2018.

[9] D. Wu, V. J. Lawhern, S. Gordon, B. J. Lance, and C.-T. Lin, "Driver drowsiness estimation from EEG signals using online weighted adaptation regularization for regression (OwARR)," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1522–1535, Dec. 2017.

[10] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 1, pp. 4–19, Mar. 2022.

[11] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, Banff, AB, Canada, Apr. 2014.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015.

[13] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli, "Is feature selection secure against training data poisoning?" in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, Jul. 2015, pp. 1689–1698.

[14] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath, "Deep learning models for electrocardiograms are susceptible to adversarial attack," *Nature Med.*, vol. 26, no. 3, pp. 360–363, Mar. 2020.

[15] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Mach. Intell.*, vol. 2, no. 6, pp. 305–311, Jun. 2020.

[16] A. Bar et al., "The vulnerability of semantic segmentation networks to adversarial attacks in autonomous driving: Enhancing extensive environment sensing," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 42–52, Jan. 2021.

[17] X. Zhang and D. Wu, "On the vulnerability of CNN classifiers in EEG-based BCIs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 5, pp. 814–825, May 2019.

[18] L. Meng, C.-T. Lin, T.-P. Jung, and D. Wu, "White-box target attack for EEG-based BCI regression problems," in *Proc. Int. Conf. Neural Inf. Process.*, Sydney, NSW, Australia, Dec. 2019, pp. 476–488.

[19] X. Zhang et al., "Tiny noise, big mistakes: Adversarial perturbations induce errors in brain–computer interface spellers," *Nat. Sci. Rev.*, vol. 8, no. 4, 2021, Art. no. nwaa233.

[20] Z. Liu, L. Meng, X. Zhang, W. Fang, and D. Wu, "Universal adversarial perturbations for CNN classifiers in EEG-based BCIs," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 0460a4.

[21] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *CoRR*, vol. abs/1712.05526, 2017.

[22] P. Margaux, M. Emmanuel, D. Sébastien, B. Olivier, and M. Jérémie, "Objective and subjective evaluation of online error correction during P300-based spelling," *Adv. Hum.-Comput. Interact.*, vol. 2012, pp. 1–13, Jan. 2012.

[23] M. Tangermann et al., "Review of the BCI competition IV," *Frontiers Neurosci.*, vol. 6, no. 1, p. 55, 2012.

[24] U. Hoffmann, J. M. Vesin, T. Ebrahimi, and K. Diserens, "An efficient P300-based brain–computer interface for disabled subjects," *J. Neurosci. Methods*, vol. 167, no. 1, pp. 115–125, Jan. 2008.

[25] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[26] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Mar. 2017.

[27] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "XDAWN algorithm to enhance evoked potentials: Application to brain–computer interface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, Aug. 2009.

[28] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[29] Y. Qin, P. Xu, and D. Yao, "A comparative study of different references for EEG default mode network: The use of the infinity reference," *Clin. Neurophysiol.*, vol. 121, no. 12, pp. 1981–1991, Dec. 2010.

[30] C. Carvalhaes and J. A. de Barros, "The surface Laplacian technique in EEG: Theory and methods," *Int. J. Psychophysiol.*, vol. 97, no. 3, pp. 174–188, 2015.

[31] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw, "Spatial filter selection for EEG-based communication," *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, no. 3, pp. 386–394, Sep. 1997.

[32] C. Chang, S. Hsu, L. Pion-Tonachini, and T. Jung, "Evaluation of artifact subspace reconstruction for automatic artifact components removal in multi-channel EEG recordings," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 4, pp. 1114–1121, Apr. 2020.

[33] Y. Liu, Y. Xie, and A. Srivastava, "Neural trojans," in *Proc. IEEE Int. Conf. Comput. Design (ICCD)*, Boston, MA, USA, Nov. 2017, pp. 45–48.

[34] G. S. Dhillon et al., "Stochastic activation pruning for robust adversarial defense," in *Proc. Int. Conf. Learn. Represent.*, Vancouver, BC, Canada, May 2018.

[35] X. Jiang, X. Zhang, and D. Wu, "Active learning for black-box adversarial attacks in EEG-based brain–computer interfaces," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Xiamen, China, Dec. 2019, pp. 361–368.

[36] X. Jiang, L. Meng, S. Li, and D. Wu, "Active poisoning: Efficient backdoor attacks to transfer learning based BCIs," *Sci. China Inf. Sci.*, 2023.