REVIEW

National Science Open
2: 20220023, 2023
https://doi.org/10.1360/nso/20220023

Information Sciences

Adversarial attacks and defenses in physiological computing: a systematic review

Dongrui Wu[1,2], Jiaxin Xu[1], Weili Fang[3], Yi Zhang[4], Liuqing Yang[5], Xiaodong Xu[4,*], Hanbin Luo[3,*] & Xiang Yu[6,*]

[1]*Ministry of Education Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China;*
[2]*Zhejiang Lab, Hangzhou 311121, China;*
[3]*School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China;*
[4]*College of Public Administration, Huazhong University of Science and Technology, Wuhan 430074, China;*
[5]*Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor MI 48109, USA;*
[6]*School of Management and Sino-European Institute for Intellectual Property, Huazhong University of Science and Technology, Wuhan 430074, China*

*Corresponding authors (emails: xiaodong-xu@hust.edu.cn (Xiaodong Xu); luohbcem@hust.edu.cn (Hanbin Luo); yuxiang@hust.edu.cn (Xiang Yu))

Received 24 December 2021; Revised 10 May 2022; Accepted 26 May 2022; Published online 26 August 2022

Abstract: Physiological computing uses human physiological data as system inputs in real time. It includes, or significantly overlaps with, brain-computer interfaces, affective computing, adaptive automation, health informatics, and physiological signal based biometrics. Physiological computing increases the communication bandwidth from the user to the computer, but is also subject to various types of adversarial attacks, in which the attacker deliberately manipulates the training and/or test examples to hijack the machine learning algorithm output, leading to possible user confusion, frustration, injury, or even death. However, the vulnerability of physiological computing systems has not been paid enough attention to, and there does not exist a comprehensive review on adversarial attacks to them. This study fills this gap, by providing a systematic review on the main research areas of physiological computing, different types of adversarial attacks and their applications to physiological computing, and the corresponding defense strategies. We hope this review will attract more research interests on the vulnerability of physiological computing systems, and more importantly, defense strategies to make them more secure.

Keywords: physiological computing, brain-computer interfaces, health informatics, biometrics, machine learning, adversarial attack

Introduction

Keyboards and mouses, and recently also touchscreens, are the most popular means that a user sends commands to a computer. However, they convey little information about the psychological state of the user, e.g., cognitions, motivations and emotions, which are also very important in the development of 'smart' technology [1]. For example, on emotions, Marvin Minsky, a pioneer in artificial intelligence, pointed out early in the 1980s that [2] "the question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions."

Physiological computing [3] is "the use of human physiological data as system inputs in real time." It opens up bandwidth within human-computer interaction by enabling an additional channel of communication from the user to the computer [1], which is necessary in adaptive and collaborative human-computer symbiosis.

Common physiological data in physiological computing include the electroencephalogram (EEG), electrocorticogram (ECoG), electrocardiogram (ECG), electrooculogram (EOG), electromyogram (EMG), eye movement, blood pressure (BP), electrodermal activity (EDA), respiration (RSP), skin temperature, etc., which are recordings or measures produced by the physiological process of human beings. Their typical measurement locations are shown in Figure 1.

These signals have been widely studied in the literature in various applications, including clinic diagnostics, and wearable devices for health monitoring and human-machine interactions, as indicated by the number of publications in Table 1. The top four most frequently studied physiological signals are blood pressure, EEG, ECG, and respiration. Blood pressure and respiration are both vital signs. Multi-lead ECG has been widely used in hospitals for screening and diagnosis of cardiovascular diseases, and single-lead ECG has been incorporated into millions of smart watches and wristbands for fitness tracking and atrial fibrillation early warning [4]. EEG is popular maybe because it is the most frequently used input signal in brain-computer interfaces (BCIs) [5] for neural rehabilitation [6], consciousness evaluation [7], emotion regularization [8], text input [9], external device control [10], etc., and gold standard in certain clinic practices, e.g., seizure diagnostics [11].

Physiological signals are usually single-channel or multi-channel time series, as shown in Figure 2. In many clinical applications, the recording may last hours, days, or even longer. For example, long-term video-EEG monitoring for seizure diagnostics may need 24 hours, and ECG monitoring in intensive care units (ICUs) may last days or weeks. Wearable ECG monitoring devices, e.g., iRythm Zio Patch, AliveCor KardiaMobile, Apple Watch, and Huawei Band, are being used by millions of users. A huge amount of physiological signals are collected during these processes. Manually labeling them is very labor-intensive, and even impossible for wearable devices, given the huge number of users.

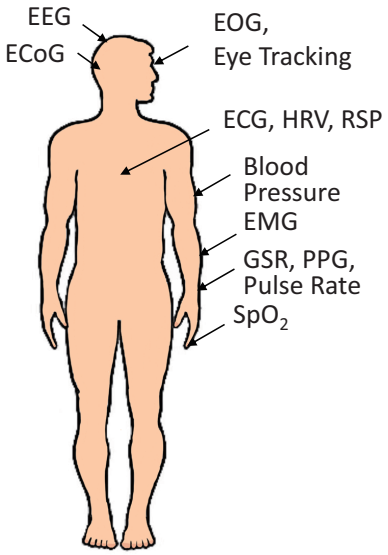

Figure 1 Common signals in physiological computing and their typical measurement locations.

Table 1 Common signals in physiological computing and the corresponding number of publications in Google Scholar with the keywords in title (as of 12/24/2021)

Signal	Keywords	No. Publications
Electro-encephalogram	EEG or electroencephalogram or electroencephalography	155,000
Electrocardiogram	ECG or EKG or electrocardiogram	104,000
Electromyogram	EMG or electromyogram	44,900
Electrocorticogram	ECoG or electrocorticogram or electrocorticography	4,340
Electrooculogram	EOG or electrooculogram	2,690
Respiration	Respiration	92,300
Blood pressure	Blood pressure	217,000
Heart rate variability	HRV or heart rate variability	40,400
Electrodermal activity	EDA or GSR or EDR or electrodermal or galvanic skin response	17,800
Eye movement	Eye movement or eye tracking	16,900
Oxygen saturation	SpO_2 or oxygen saturation or blood oxygen	26,800
Skin temperature	Skin temperature	7,320
Photo-plethysmogram	PPG or photoplethysmogram	5,390
Pulse rate	Pulse rate	7,160

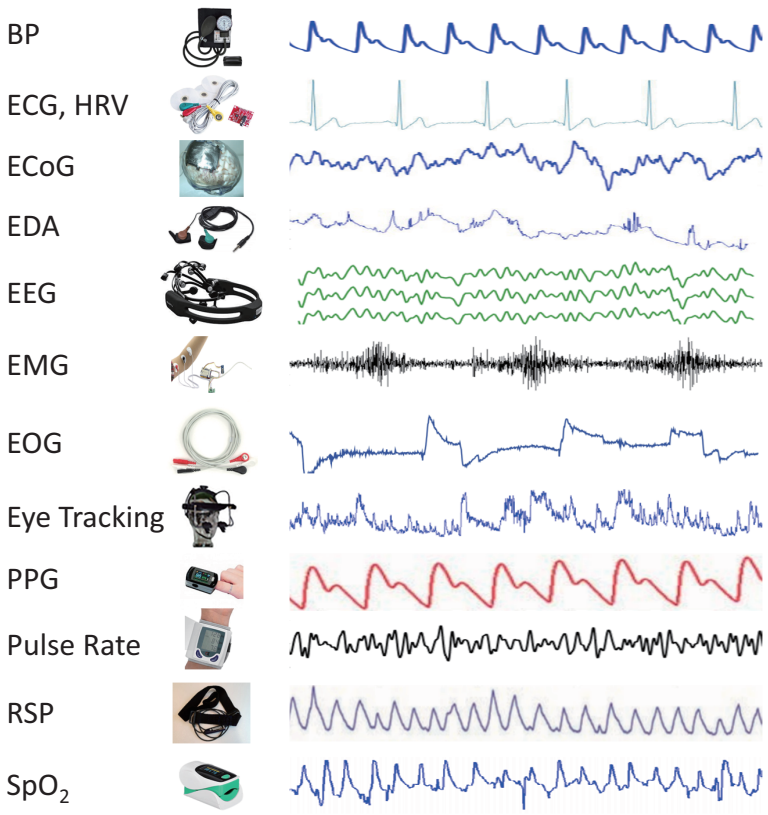

Figure 2 Examples of common signals in physiological computing, and their typical measurement equipment. Blood pressure is usually measured by auscultation. ECG/HRV, ECoG, EDA, EEG, EMG, and EOG are bioelectrical signals generated by nerves and muscles. Eye tracking, PPG and SpO_2 are bio-optical signals.

Machine learning [12] has been used to alleviate this problem, by automatically classifying the measured physiological signals. Particularly, deep learning has demonstrated outstanding performances [13], e.g., EEGNet [14], DeepCNN [15], ShallowCNN [15], and TIDNet [16] for EEG classification, SeizureNet for EEG-based seizure recognition [17], CNN for ECG rhythm classification [18], ECGNet for ECG-based mental stress monitoring [19], and so on.

However, recent research has shown that both traditional machine learning and deep learning models are vulnerable to various types of attacks [20–23]. For example, Sharif *et al.* [24] successfully fooled a face recognition system using a deliberately designed eyeglass for the target face. Brown *et al.* [25] generated adversarial patches, which could be placed anywhere within an image and caused the classifier to output the target class. Chen *et al.* [26] created a backdoor in the target model by injecting poisoning samples, which contain an ordinary sunglass, into the training set, so that all test images with the sunglass would be classified into a target class. Athalye *et al.* [27] synthesized a 3D adversarial turtle, which was classified as a rifle at every viewpoint. Eykholt *et al.* [28] stuck a carefully crafted graffiti to road signs, and caused the model to classify 'Stop' as 'Speed limit 40'. Finlayson *et al.* [29, 30] successfully performed adversarial attacks to deep learning classifiers across three clinical domains (fundoscopy, chest X-ray, and dermoscopy). Rahman *et al.* [31] performed adversarial attacks to six COVID-19 related applications, including recognizing whether a subject is wearing a mask, maintaining deep learning based QR codes as immunization certificates, and recognizing COVID-19 from CT scan or X-ray images. Ma *et al.* [32] showed that medical deep learning models can be more vulnerable to adversarial attacks than models for natural images, but surprisingly and fortunately, medical adversarial attacks may also be easily detected. Kaissis *et al.* [33] pointed out that various other attacks, in addition to adversarial attacks, also exist in medical imaging, and called for secure, privacy-preserving and federated machine learning to cope with them.

Machine learning models in physiological computing are not exempt from adversarial attacks [34–36]. However, to the best of our knowledge, there does not exist a systematic review on adversarial attacks in physiological computing. This study fills this gap, by comprehensively reviewing different types of adversarial attacks, their applications in physiological computing, and possible defense strategies. It will be very important to the security of physiological computing systems in real-world applications.

We need to emphasize that this study focuses on the emerging adversarial attacks and defenses. For other types of attacks and defenses, e.g., cybersecurity, the readers can refer to, e.g., [37].

The remainder of this study is organized as follows. Section 2 introduces five relevant research areas in physiological computing. Section 3 introduces different categorizations of adversarial attacks. Section 4 describes various adversarial attacks to physiological computing systems. Section 5 introduces different approaches to defend against adversarial attacks, and their applications in physiological computing. Finally, Section 6 draws conclusions and points out some future research directions.

Physiological computing

As physiological computing uses human physiological data as system inputs in real time, it includes, or significantly overlaps with, BCIs, affective computing, adaptive automation, health informatics, and physiological signal based biometrics. Although these five areas have different application scenarios and goals, they

all need to build machine learning models for physiological signal classification or regression, and hence all are subject to adversarial attacks.

BCIs

A BCI system establishes a direct communication pathway between the brain and an external device, e.g., a computer or a robot [5]. Scalp and intracranial EEGs have been widely used in BCIs [12].

The flowchart of a closed-loop EEG-based BCI system is shown in Figure 3. After EEG signal acquisition, signal processing, usually including both temporal filtering and spatial filtering, is used to enhance the signal-to-noise ratio. Machine learning is next performed to understand what the EEG signal means, based on which a control command may be sent to an external device.

There are three typical paradigms in EEG-based BCIs [12].

(1) Motor imagery (MI) [38], which modifies neuronal activities in primary sensorimotor areas when the user imagines the movement of various body parts, e.g., the left (right) hemisphere for right-hand (left-hand) MIs and center for feet MIs. These MIs can be decoded to control external devices, e.g., a wheelchair, or in neural rehabilitation [6] to restore the functionality of hands after stroke.

(2) Event-related potentials (ERPs) [39, 40], which are stereotyped EEG responses to rare or expected visual, audio, or tactile stimuli. The most frequently used ERP component is P300 [41], which is an increase of the EEG magnitude observed about 300 ms after a rare stimulus.

(3) Steady-state visual evoked potential (SSVEP) [42], which is brain's electrical response to repetitive visual stimulation, usually between 3.5 and 75 Hz [43]. SSVEP can achieve very high information transfer rate in BCI spellers [9].

EEG-based BCI spellers may be the only non-muscular communication devices for amyotrophic lateral sclerosis (ALS) patients to express their opinions [44]. In seizure treatment, responsive neurostimulation (RNS) [45, 46] recognizes ECoG or intracranial EEG patterns prior to ictal onset, and delivers a high-frequency stimulation impulse to stop the seizure, improving the patient's quality-of-life.

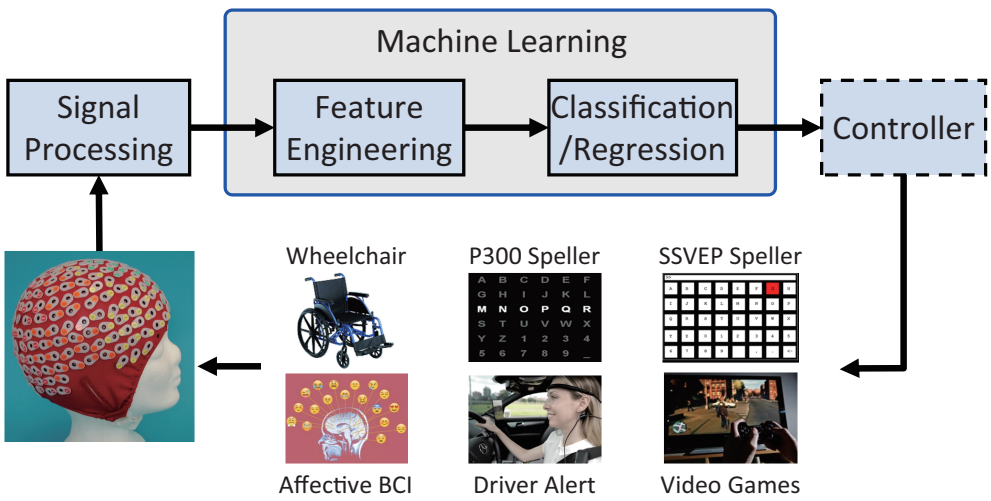

Figure 3 Flowchart of a closed-loop EEG-based BCI system.

Affective computing

Affective computing [47] is "computing that relates to, arises from, or deliberately influences emotion or other affective phenomena."

Emotion is the focus of affective computing. It can be represented by discrete categories, e.g., Ekman's six basic emotions [48] (anger, disgust, fear, happiness, sadness, and surprise), and by continuous values in the 2D space of arousal and pleasure (or valence) [49], or the 3D space of arousal, pleasure (or valence) and dominance [50], as shown in Figure 4.

Various inputs can be used in affective computing, e.g., videos, text, and speech. Physiological signals have also been extensively used [51]. In bio-feedback based relaxation training [52], EDA can be used to detect the user's affective state, based on which a relaxation training application can provide the user with explicit feedbacks to learn how to change his/her physiological activities to improve health and performance. In software adaptation [53], the graphical interface, difficulty level, sound effects and/or contents are automatically adapted based on the user's real-time emotion estimated from various physiological signals, to keep the user more engaged.

Adaptive automation

Adaptive automation controls the number and/or types of tasks allocated to the operator to keep the workload within an appropriate level (avoiding both underload and overload), and hence to enhance the overall performance and safety of the human-machine system [54, 55].

Boeke *et al.* [54] expressed adaptive automation as a control system, as shown in Figure 5. The task demand estimator maps each task to a cognitive demand. The dynamic task allocator allocates tasks to the operator based upon his/her available cognitive capacity and the incoming tasks' cognitive demands. The available cognitive capacity estimator estimates the operator's available cognitive capacity from his/her performance, physiology, or subjective measures.

In air traffic management [55], an operator's EEG signal can be used to estimate the mental workload, and trigger specific adaptive automation solutions. This can significantly reduce the operator's workload during high-demanding conditions, and increase the task execution performance. Ref. [56] also showed that pupil diameter and fixation time, measured from an eye-tracking device, can be indicators of mental workload, and

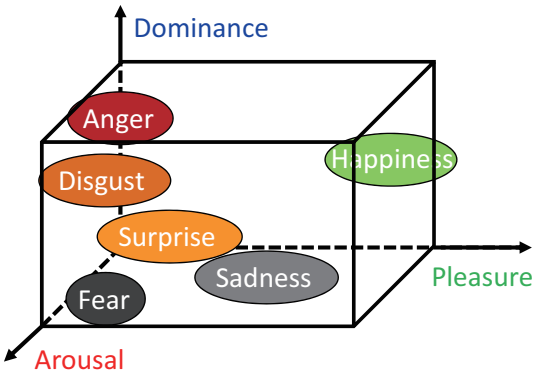

Figure 4 Ekman's six basic emotions in the 3D space of arousal, pleasure and dominance.

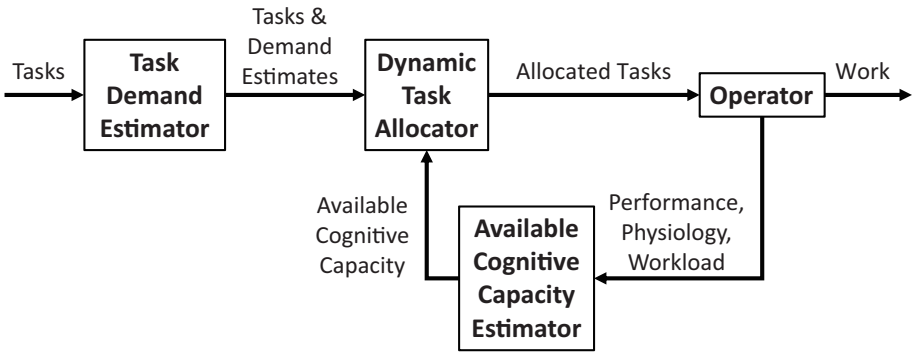

Figure 5 Adaptive automation expressed as a control system [54].

hence be used to trigger adaptive automation.

Park and Zahabi [57] performed a review on cognitive workload assessment of prosthetic devices, which can be achieved using physiological, subjective, or task performance measures. The first includes EEG, EMG, ECG, EDA, respiration, and eye-tracking. They found that hybrid inputs, e.g., EMG plus inertial measurement unit (IMU), or EMG plus force myography (FMG), were less cognitively demanding than EMG or EEG alone. More specifically, the combination of EMG and IMU can improve the effectiveness, satisfaction and efficiency of prosthetic devices, and the combination of EMG and FMG achieved higher overall stability (lower variance) than EMG alone.

Health informatics

Health informatics studies information and communication processes and systems in healthcare [58].

A single-lead short ECG recording (9–60 s), collected from the AliveCor personal ECG monitor, can be used by a convolutional neural network (CNN) to classify normal sinus rhythm, atrial fibrillation, an alternative rhythm, or noise, with an average test accuracy of 88% on the first three classes [4]. Ref. [59] also showed that heart rate data from consumer smart watches, e.g., Apple, Fitbits and Garmin devices, can be used for pre-symptomatic detection of COVID-19, sometimes nine or more days earlier.

Many smart watches or wristbands record the PPG signal, a measure of arterial blood volume fluctuating with each heartbeat [60]. It is frequently used to monitor the heart rate, but also contains information on the cardiac, vascular, respiratory, and autonomic nervous systems. In clinics, wearable PPGs can be used for atrial fibrillation detection, obstructive sleep apnea identification, monitoring the spread of infectious diseases, sleep monitoring, mental stress assessment, vascular age assessment, clinical deterioration identification, cardiovascular risk prediction, response to exercise assessment, sepsis identification, heart failure identification, and preeclampsia identification [60]. For example, Guo *et al.* [61] proposed a model fusion approach (Figure 6) for PPG-based atrial fibrillation onset prediction, using Huawei smart watches and wristbands. It achieved 94.04% sensitivity, 96.35% specificity, and 94.04% recall.

Physiological signal based biometrics

Physiological signal based biometrics [62] use physiological signals for biometric applications, e.g., digitally identifying a person to grant access to systems, devices or data. They represent a paradigm shift from

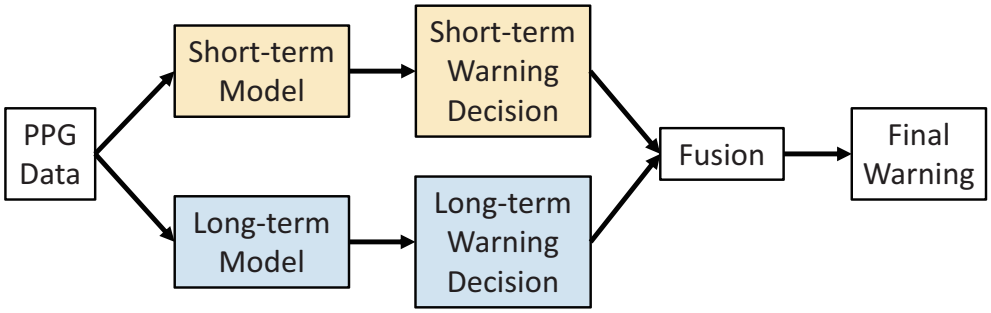

Figure 6 A model fusion approach for PPG-based atrial fibrillation onset prediction [61].

conventional "something we know" (e.g., a personal identification number) or "something we have" (e.g., an access card) policies to "something we are" [63].

A biometric system typically includes three modes [63]: enrollment, identification, and authentication, as shown in Figure 7. The enrollment mode converts each subject's physiological signals into a feature template and stores it in a database. The identification mode finds the best possible match for an incoming subject's template. The authentication mode verifies if a subject is indeed the person he/she claims to be.

EEG [63], ECG [64], PPG [65], and multimodal physiological signals [66] have been used in user identification and authentication, with the advantages of universality, permanence, liveness detection, continuous authentication, etc. Thomas and Vinod [63] performed a review on EEG-based biometric systems, and found that EEGs from the resting state with eyes closed or open, motor imagination, visual evoked potentials, and mental tasks (e.g., math operation, letter composition) can all be used in biometrics. Agrafioti *et al.* [64] gave a comprehensive introduction of the theory, methods and applications of heart biometrics, pointing out their challenges, including time dependency, collection periods, privacy implications, and cardiac conditions. Bianco and Napoletano [66] performed multimodal biometric recognition using heart rate, breathing rate, palm EDA, and perinasal perspiration, achieving 90.54% top-1 accuracy and 99.69% top-5 accuracy.

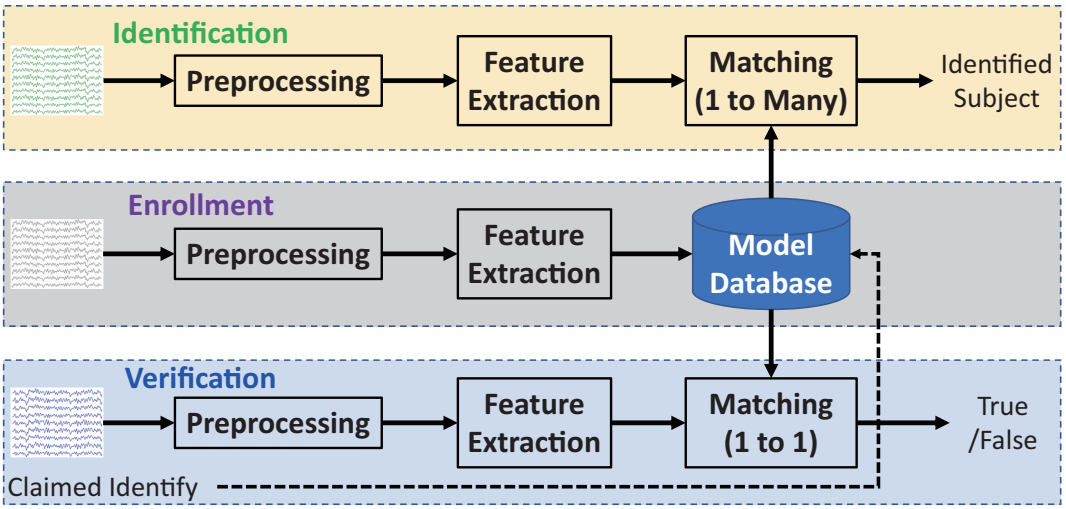

Figure 7 Enrollment, identification and authentication in biometrics [63].

Adversarial attacks

Adversarial attacks generate various adversarial perturbations, which may be unnoticeable by human eyes or a computer program, to fool a machine learning model. There are different categorizations of adversarial attacks [22, 23], as shown in Figure 8.

Targeted and non-targeted attacks

According to the outcome, there are two types of adversarial attacks [23]: targeted attacks and non-targeted (indiscriminate) attacks.

Targeted attacks force a model to classify certain examples, or a certain region of the feature space, into a specific (usually wrong) class. Non-targeted attacks force a model to misclassify certain examples or feature space regions, but do not specify which class they should be misclassified into.

For example, in a 3-class classification problem, assume the class labels are A, B and C. Then, a targeted attack may force the input to be classified into class A, no matter what its true class is. A non-targeted attack forces an input from class A to be classified into classes B or C, but does not specify it must be B or C; as long as it is not A, then the non-targeted attack is successful.

White-box, black-box and gray-box attacks

According to how much the attacker knows about the target model, there can be three types of attacks [67].

(1) White-box attacks, in which the attacker knows everything about the target model, including its architecture and parameters. This is the easiest attack scenario and could cause the maximum damage. It may

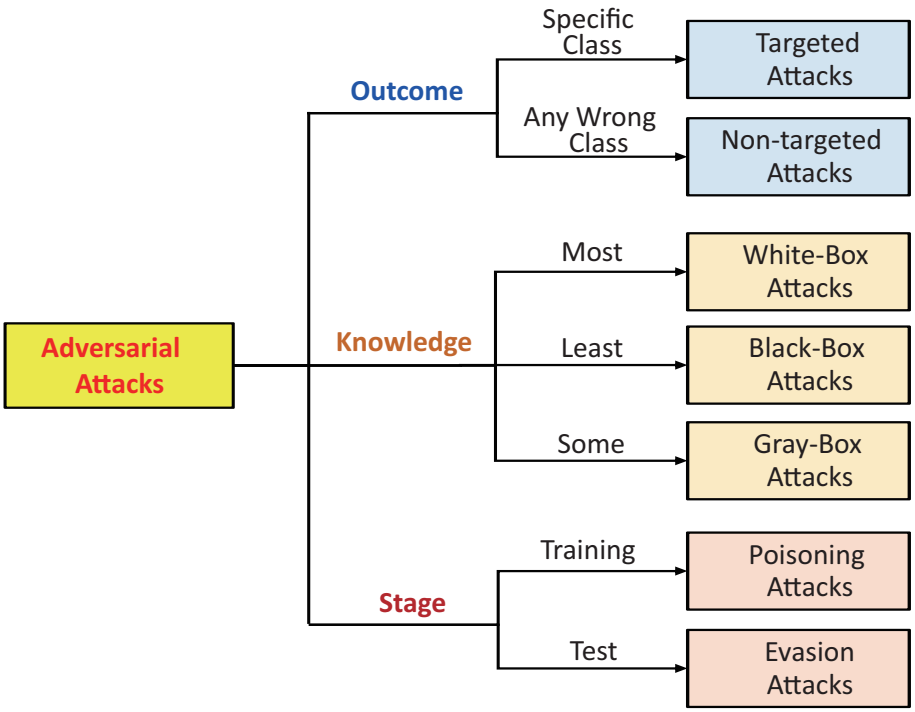

Figure 8 Types of adversarial attacks.

correspond to the case that the attacker is an insider, or the model designer is evaluating the worst-case scenario when the model is under attack. Popular attack approaches include L-BFGS [20], DeepFool [68], the C&W method [69], the fast gradient sign method (FGSM) [21], and the basic iterative method (BIM) [70].

(2) Black-box attacks, in which the attacker knows neither the architecture nor the parameters of the target model, but can supply inputs to the model and observe its outputs. This is the most realistic and also the most challenging attack scenario. One example is that the attacker purchases a commercial BCI system and tries to attack it. Black-box attacks are possible, due to the transferability of adversarial examples [20]; i.e., an adversarial example generated from one machine learning model may be used to fool another machine learning model at a high success rate, if the two models solve the same task. So, in black-box attacks [71], the attacker can query the target model many times to construct a training set, train a substitute machine learning model from it, and then generate adversarial examples from the substitute model to attack the original target model.

(3) Gray-box attacks, which assume the attacker knows a limited amount of information about the target model, e.g., (part of) the training data that the target model is tuned on. They are frequently used in data poisoning attacks, as introduced in the next subsection.

Table 2 compares the main characteristics of the three attack types.

Poisoning and evasion attacks

According to the stage that the adversarial attack is performed, there are two types of attacks: poisoning attacks and evasion attacks, as shown in Figure 9.

Poisoning attacks [72] focus on the training stage, to create backdoors in the machine learning model by adding contaminated examples to the training set. At the test stage, an input with the backdoor can be classified into the class the attacker specifies. They are usually white-box or gray-box attacks, achieved by data injection, i.e., adding adversarial examples to the training set [73], or data modification, i.e., poisoning the training data by modifying their features or labels [74].

Evasion attacks [21] happen at the test stage, by adding deliberately designed tiny perturbations to benign test examples to mislead the machine learning model. They are usually white-box or black-box attacks.

Adversarial attacks in physiological computing

Most adversarial attack studies considered computer vision applications, where the inputs are 2D images. Physiological signals are continuous time series, which are quite different from images. There are relatively

Table 2 Comparison of white-box, gray-box and black-box attacks [67]. '–' means that whether the information is available or not does not affect the attack strategy, since it is not used in the attack.

Target model information	White-box	Gray-box	Black-box
Know its architecture	✓	×	×
Know its parameters	✓	×	×
Know its training data	–	✓	×
Can observe its response	–	–	✓

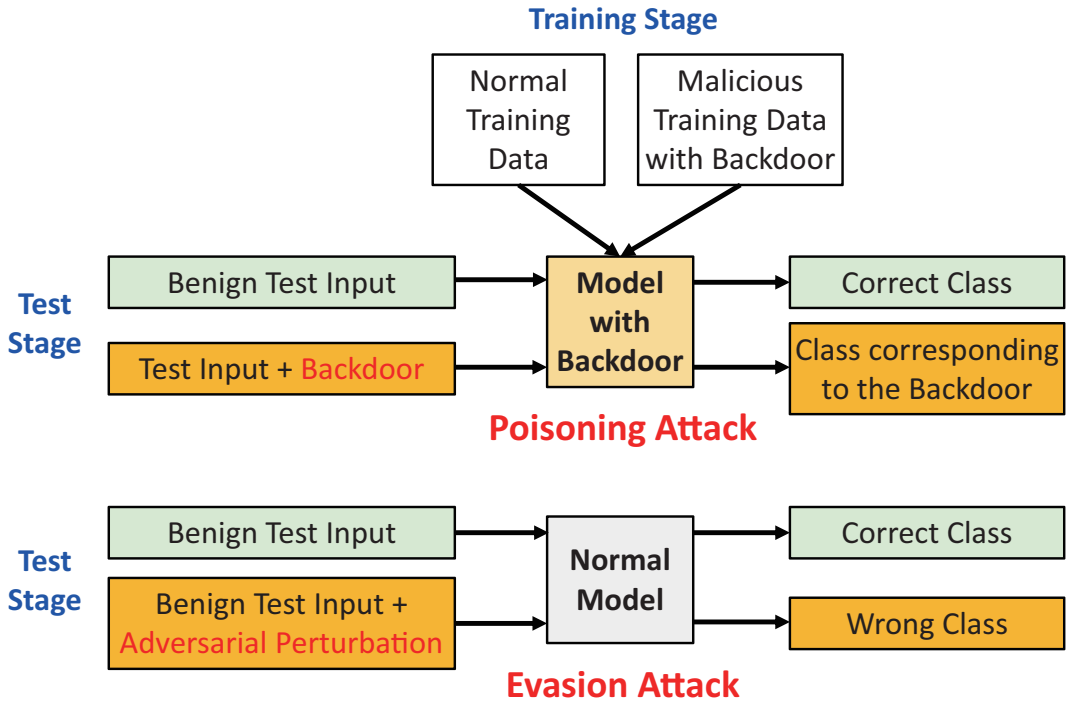

Figure 9 Poisoning and evasion attacks.

few adversarial attack studies on time series [75–78], and even fewer on physiological signals. A summary of them is shown in Table 3 [4, 34–36, 67, 76, 79–88].

Adversarial attacks in BCIs

Attacking the machine learning models in BCIs could cause significant damages, ranging from user frustration to serious injuries. For example, in seizure treatment, attacks to RNS's [45] seizure recognition algorithm may quickly drain its battery or make it completely ineffective, significantly reducing the patient's quality-of-life. Adversarial attacks to an EEG-based BCI speller may hijack the user's true inputs and output wrong letters, leading to user frustration or misunderstanding. In BCI-based driver drowsiness estimation [89], adversarial attacks may make a drowsy driver look alert, increasing the risk of accidents.

Although most BCI research so far focused on making BCIs faster and more accurate, pioneers in BCIs have started to consider neurosecurity. For example, Ienca *et al.* [90] pointed in a *Nature Biotechnology* Commentary in 2018 that "greater safeguards are needed to address the personal safety, security and privacy risks arising from increasing adoption of neurotechnology in the consumer realm." Jarchum [91] pointed out in a *Nature Biotechnology* Focus article in 2019 three concerns on the widespread use of brain recording and stimulation. The second is "devices getting hacked and, by extension, behavior unwillfully and unknowingly manipulated for nefarious purposes (although this could conceivably lead to checked bad behavior too)." A 2020 RAND Corporation report [92] pointed out that "hacking BCI capabilities could theoretically provide adversaries with direct pathways into the emotional and cognitive centers of operators' brains to sow confusion or emotional distress. In the extreme, adversary hacking into BCI devices that influence the motor cortex of human operators could theoretically send false directions or elicit unintended actions, such as friendly fire,

Table 3 Summary of existing adversarial attack approaches in physiological computing

Application	Problem	Ref.	Outcome		Knowledge			Stage	
			Targeted	Non-targeted	White-box	Gray-box	Black-box	Poisoning	Evasion
BCI	Classification	[67]		✓	✓	✓	✓		✓
		[79]		✓			✓		✓
		[34]	✓		✓				✓
		[80]	✓	✓	✓	✓			✓
		[81]	✓				✓	✓	
		[82]	✓			✓			✓
	Regression	[83]	✓		✓		✓		✓
Health informatics	Classification	[4]	✓	✓	✓				✓
		[84]	✓		✓				✓
		[85]	✓	✓	✓		✓	✓	✓
		[86]	✓		✓			✓	
		[76]		✓	✓		✓		✓
Biometrics	Classification	[87]	✓				✓		✓
		[88]	✓			✓			✓
		[36]	✓			✓			✓
		[35]	✓			✓			✓

although such influence may be technically difficult to achieve in the near term. Even an attack that broadly degraded gross motor skills could prove debilitating during combat." In fact, as introduced below, adversarial attacks to EEG-based BCIs have become more and more practical.

In 2019, Zhang and Wu [67] first pointed out that adversarial examples exist in EEG-based BCIs, i.e., deep learning models in BCIs are vulnerable to adversarial attacks. They successfully performed white-box, gray-box and black-box non-targeted evasion attacks to three CNN classifiers, i.e., EEGNet [14], DeepCNN and ShallowCNN [15], in three different BCI paradigms, i.e., P300 evoked potential detection, feedback error-related negativity detection, and motor imagery classification. The basic idea, shown in Figure 10, is to add a jamming module between EEG signal processing and machine learning to generate adversarial examples, optimized by unsupervised FGSM. The generated adversarial perturbations are too small to be noticed by human eyes (an example is shown in Figure 11), but can significantly reduce the classification accuracy.

It is important to note that the jamming module is implementable, as Ref. [93] has shown that BtleJuice, a framework to perform Man-in-the-Middle attacks on Bluetooth devices, can be used to intercept the data

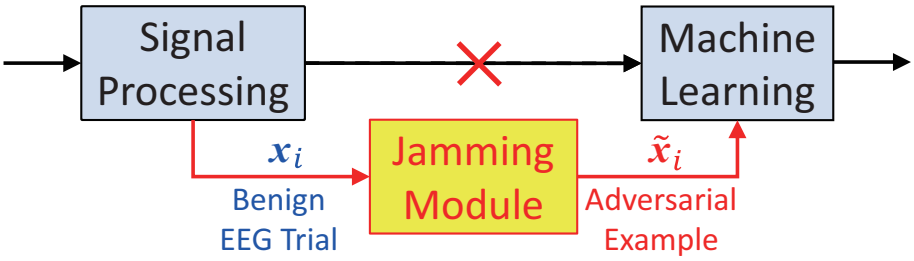

Figure 10 The BCI evasion attack approach proposed in [67]. A jamming module is inserted between signal preprocessing and machine learning to generate adversarial examples.

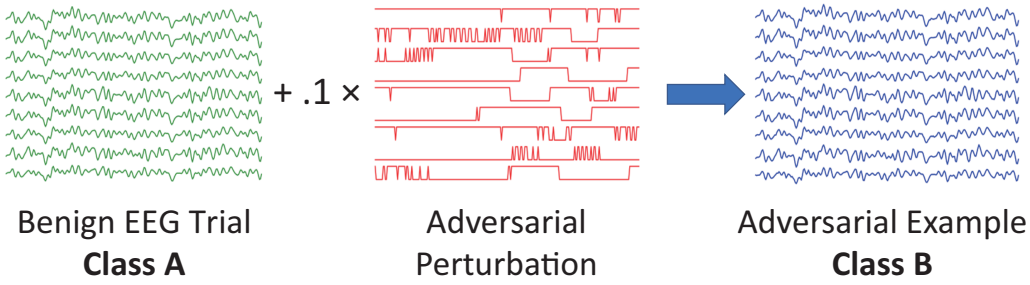

Figure 11 Evasion attack in BCIs [67].

from a consumer grade EEG-based BCI system, modify them, and then send them back to the headset. The RAND report [92] also pointed out that "in a battlefield situation, these weak signals (electrical signals in BCIs) could potentially be jammed."

Jiang *et al.* [79] focused on black-box non-targeted evasion attacks to deep learning models in BCI classification problems, in which the attacker trains a substitute model to approximate the target model, and then generates adversarial examples from the substitute model to attack the target model. Learning a good substitute model is critical to the success of black-box attacks, but it requires a large number of queries to the target model. Jiang *et al.* [79] proposed a novel query synthesis based active learning framework to improve the query efficiency, by actively synthesizing EEG trials scattering around the decision boundary of the target model, as shown in Figure 12. Compared with the original black-box attack approach in [67], the active learning based approach can improve the attack success rate with the same number of queries, or, equivalently, reduce the number of queries to achieve a desired attack performance. This is the first work that integrates active learning and adversarial attacks for EEG-based BCIs.

The above two studies considered classification problems, as in most adversarial attack research. Adversarial attacks to regression problems were much less investigated in the literature. Meng *et al.* [83] were the first to study white-box targeted evasion attacks for BCI regression problems. They proposed two approaches, based on optimization and gradient, respectively, to design small perturbations to change the regression output by a pre-determined amount. Experiments on two BCI regression problems (EEG-based driver fatigue estimation, and EEG-based user reaction time estimation in the psychomotor vigilance task) verified their effectiveness: both approaches can craft adversarial EEG trials indistinguishable from the original ones, but can significantly change the outputs of the BCI regression model. Moreover, adversarial examples generated from both approaches are also transferable; i.e., adversarial examples generated from one known regression

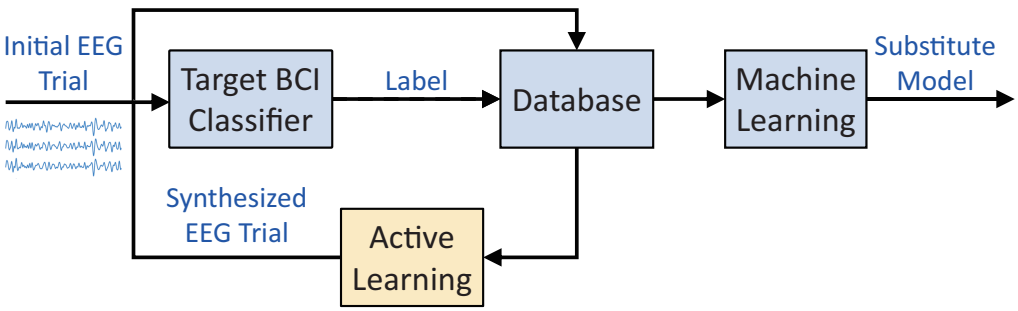

Figure 12 Query synthesis based active learning in black-box evasion attack to EEG-based BCIs [79].

model can also be used to attack an unknown regression model in black-box settings.

The above three attack strategies are theoretically important, but there are some constraints in applying them to real-world BCIs.

(1) Trial-specificity, i.e., the attacker needs to generate different adversarial perturbations for different EEG trials.

(2) Channel-specificity, i.e., the attacker needs to generate different adversarial perturbations for different EEG channels.

(3) Non-causality, i.e., the complete EEG trial needs to be known in advance to compute the corresponding adversarial perturbation.

(4) Synchronization, i.e., the exact starting time of the EEG trial needs to be known for the best attack performance.

Some recent studies tried to overcome these constraints.

Zhang *et al.* [34] performed white-box targeted evasion attacks to P300 and SSVEP based BCI spellers (Figure 13), and showed that a tiny perturbation to the EEG trial can mislead the speller to output any character the attacker wants, e.g., change the output from 'Y' to 'N', or vice versa. The most distinguishing characteristic of their approach is that it explicitly considers the causality in designing the perturbation; i.e., it should be generated before or as soon as the target EEG trial starts, so that it can be added to the EEG trial in real-time in practice. To achieve this, an adversarial perturbation template is constructed from the training set only and then fixed. So, there is no need to know the test EEG trial and compute the perturbation specifically for it. Their approach resolves the trial-specificity and non-causality constraints, but different EEG channels still need different perturbations, and it also requires the attacker to know the starting time of an EEG trial in advance to achieve the best attack performance, i.e., there are still channel-specificity and synchronization constraints.

Zhang *et al.* [34] considered targeted attacks to a traditional and most frequently used BCI speller pipeline, which has separate feature extraction and classification steps. Liu *et al.* [80] considered both targeted and non-targeted white-box evasion attacks to end-to-end deep learning models in EEG-based BCIs, and pro-

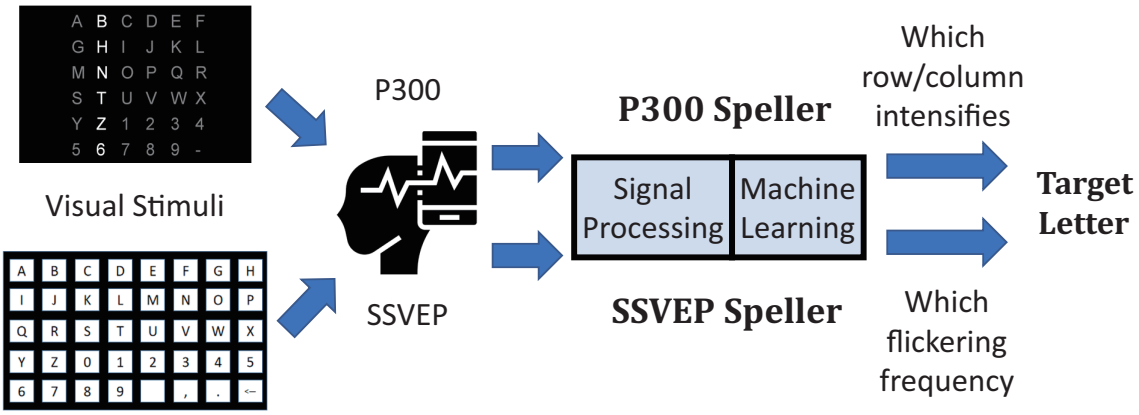

Figure 13 Workflow of a P300 speller and an SSVEP speller [34]. For each speller, the user watches the stimulation interface, focusing on the character he/she wants to input, while EEG signals are recorded and analyzed by the speller. The P300 speller first identifies the row and the column that elicit the largest P300, and then outputs the letter at their intersection. The SSVEP speller identifies the output letter directly by matching the user's EEG oscillation frequency with the flickering frequency of each candidate letter.

posed a total loss minimization (TLM) approach to generate universal adversarial perturbations (UAPs) for them. Experimental results demonstrated its effectiveness on three popular CNN classifiers (EEGNet, ShallowCNN, and DeepCNN) in three BCI paradigms (P300, feedback error related negativity, and motor imagery). They also verified the transferability of UAPs in non-targeted gray-box evasion attacks.

To further simplify the implementation of TLM-UAP, Liu *et al.* [80] also considered smaller template size, i.e., mini TLM-UAP with a small number of channels and time domain samples, which can be added anywhere to an EEG trial. Mini TLM-UAPs are more practical and flexible, because they do not require the attacker to know the exact number of EEG channels and the exact length and starting time of an EEG trial. Liu *et al.* [80] showed that, generally, all mini TLM-UAPs were effective. However, their effectiveness decreased when the number of used channels and/or the template length decrease, which is intuitive. This is the first study on UAPs of CNN classifiers in EEG-based BCIs, and also the first on optimization based UAPs for targeted evasion attacks.

In summary, the TLM-UAP approach [80] resolves the trial-specificity and non-causality constraints, and mini TLM-UAPs further alleviate the channel-specificity and synchronization constraints.

More recently, Bian *et al.* [82] proposed square wave evasion attacks to two popular training-free models (canonical correlation analysis, and filter-bank canonical correlation analysis) in SSVEP-based BCI spellers. As shown in Figure 14, the attacker only needs to know the frequency of the target character; by adding a square wave of that frequency to any input SSVEP trial, the output character can be changed to the target character with almost 100% success rate. The proposed attack approach can resist EEG preprocessing, is robust to SSVEP trial length, and is insensitive to the phase of the square wave signal; i.e., the attacker can use any random initial phase. This represents so far the easiest implementation of evasion attacks to SSVEP-based BCI systems.

All above studies focused on evasion attacks. Meng *et al.* [81] were the first to show that poisoning attacks can also be performed for EEG-based BCIs, as shown in Figure 15. They proposed a practically realizable backdoor key, narrow period pulse, for EEG signals, which can be inserted into the benign EEG signal during data acquisition, and demonstrated its effectiveness in black-box targeted poisoning attacks, i.e., the attacker does not know any information about the test EEG trial, including its starting time, and wants to classify the test trial into a specific class, regardless of its true class. In other words, it resolves the trial-specificity, channel-specificity, causality and synchronization constraints simultaneously. To our knowledge, this is to-date the most practical BCI attack approach.

A summary of existing adversarial attack approaches in EEG-based BCIs is shown in Table 4.

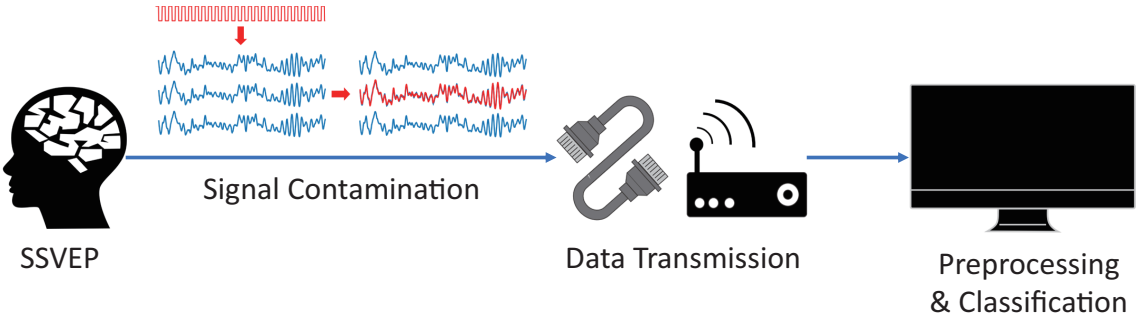

Figure 14 Square wave evasion attack to SSVEP spellers [82].

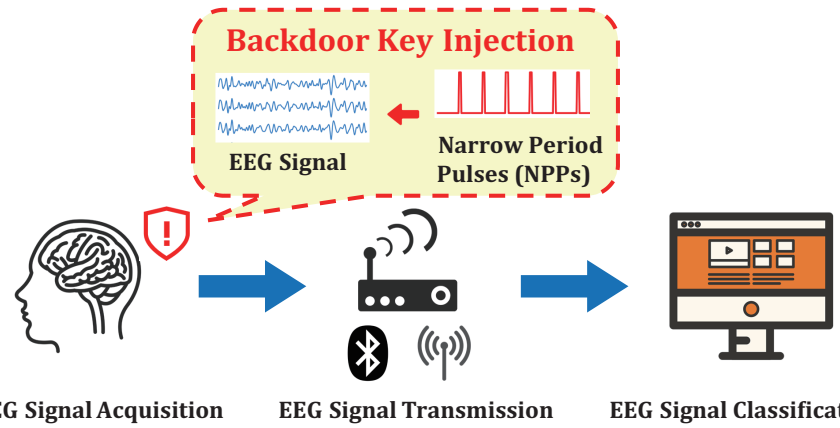

Figure 15 Poisoning attack in EEG-based BCIs [81]. Narrow period pulses can be added to EEG trials during signal acquisition.

Table 4 Characteristics of existing adversarial attack approaches in EEG-based BCIs. '✓' means that constraint is satisfied, '~' is partially resolved, and '×' is not satisfied.

Ref.	Trial-specificity	Channel-specificity	Non-causality	Synchronization
[67]	×	×	×	×
[79]	×	×	×	×
[83]	×	×	×	×
[34]	✓	×	✓	×
[80]	✓	~	✓	~
[81]	✓	✓	✓	✓
[82]	✓	✓	✓	✓

Adversarial attacks in health informatics

Adversarial attacks in health informatics can also cause serious damages, even deaths. For example, adversarial attacks to the machine learning algorithms in implantable cardioverter defibrillators could lead to unnecessary painful shocks, damaging the cardiac tissue, and even worse therapy interruptions and sudden cardiac death [94].

Han *et al.* [4] proposed both targeted and non-targeted white-box evasion attack approaches to construct smoothed adversarial examples for ECG trials that are invisible to one board-certified medicine specialist and one cardiac electrophysiology specialist, but can successfully fool a CNN classifier for arrhythmia detection. They achieved 74% attack success rate (74% of the test ECGs originally classified correctly were assigned a different diagnosis, after adversarial attacks) on atrial fibrillation classification from single-lead ECG collected from the AliveCor personal ECG monitor. This study suggests that it is important to check if ECGs have been altered before using them in medical machine learning models.

Aminifar [84] studied white-box targeted evasion attacks in EEG-based epileptic seizure detection, through UAPs. He computed the UAPs via solving an optimization problem, and showed that they can fool a support vector machine classifier to misclassify most seizure samples into non-seizure ones, with imperceptible amplitude.

Newaz *et al.* [85] investigated adversarial attacks to machine learning-based smart healthcare systems,

consisting of 10 vital signs, e.g., EEG, ECG, SpO_2, respiration, blood pressure, blood glucose, and blood hemoglobin. They performed both targeted and non-targeted attacks, and both poisoning and evasion attacks. For evasion attacks, they also considered both white-box and black-box attacks. They showed that adversarial attacks can significantly degrade the performance of four different classifiers in smart health system in detecting diseases and normal activities, which may lead to erroneous treatment.

Deep learning has been extensively used in health informatics; however, generally it needs a large amount of training data for satisfactory performance. Transfer learning [12] can be used to alleviate this requirement, by making use of data or machine learning models from an auxiliary domain or task. Wang *et al.* [86] studied targeted backdoor attacks against transfer learning with pre-trained deep learning models on both image and time series (e.g., ECG). Three optimization strategies, i.e., ranking-based neuron selection, autoencoder-powered trigger generation, and defense-aware retraining, were used to generate backdoors and retrain deep neural networks, to defeat pruning based, fine-tuning/retraining based and input pre-processing based defenses. They demonstrated their effectiveness in brain MRI image classification and ECG heartbeat type classification.

Adversarial attacks in biometrics

Physiological signals, e.g., EEG, ECG and PPG, have recently been used in biometrics [62]. However, they are subject to presentation attacks in such applications. In a physiological signal based presentation attack, the attacker tries to spoof the biometric sensors with a fake piece of physiological signal [88], which would be authenticated as from a specific victim user.

Maiorana *et al.* [87] investigated the vulnerability of an EEG-based biometric system to hill-climbing attacks. They assumed that the attacker can access the matching scores of the biometric system, which can then be used to guide the generation of synthetic EEG templates until a successful authentication is achieved. This is essentially a black-box targeted evasion attack in the adversarial attack terminology: the synthetic EEG signal is the adversarial example, and the victim's identify is the target class. It is a black-box attack, because the attacker can only observe the output of the biometric system, but does not know anything else about it.

Eberz *et al.* [88] proposed an offline ECG biometrics presentation attack approach, illustrated in Figure 16A. The basic idea was to find a mapping function to transform ECG trials recorded from the attacker so that they resemble the morphology of ECG trials from a specific victim. The transformed ECG trials can then be used to fool an ECG biometric system to obtain unauthorized access. They showed that the attacker ECG trials can be obtained from a device different from the one that the victim ECG trials are recorded from (i.e., cross-device attack), and there could be different approaches to present the transformed ECG trials to the biometric device under attack, the simplest being the playback of ECG trials encoded as .wav files using an off-the-shelf audio player.

Unlike [87], the above approach is a gray-box targeted evasion attack in the adversarial attack terminology: the attacker's ECG signal can be viewed as the benign example, the transformed ECG signal is the adversarial example, and the victim's identity is the target class. The mapping function plays the role of the jamming module in Figure 10. It is a gray-box attack, because the attacker needs to know the feature distributions of the victim ECGs in designing the mapping function.

Karimian *et al.* [36] proposed an online ECG biometrics presentation attack approach, shown in Figure 16A. Its procedure is very similar to the offline attack one in Figure 16B [95], except that the online approach is simpler, because it only requires as few as one victim ECG segment to compute the mapping function, and the mapping function is linear. Karimian [35] also proposed a similar presentation attack approach to attack PPG-based biometrics. Again, these approaches can be viewed as gray-box targeted evasion attacks.

More recently, Karim *et al.* [76] utilized an adversarial transformation network on a distilled model to attack two classification models (1-nearest neighbor dynamic time warping, and a fully convolutional network) on 42 time series datasets. There were three ECG datasets on the classification of humans with heart conditions or Myocardial infarctions, and two EOG datasets on the classification of Japanese Katakana strokes. They performed both white-box and black-box non-targeted evasion attacks. However, they did not consider the causality of the time-series, i.e., the entire test trial was used in generating the adversarial perturbation. So, their approach may only be used offline.

Discussion

Although we have not found adversarial attack studies on affective computing and adaptive automation in physiological computing, it does not mean that adversarial attacks cannot be performed in such applications. Machine learning models in affective computing and adaptive automation are not fundamentally different from those in BCIs; so, adversarial attacks in BCIs can easily be adapted to affective computing and adaptive automation. Particularly, Meng *et al.* [83] have shown that it is possible to attack the regression models in EEG-based driver fatigue estimation and EEG-based user reaction time estimation, whereas driver fatigue and user reaction time could be triggers in adaptive automation.

It is interesting to note that almost all aforementioned adversarial attacks focused on EEG and ECG, the 2nd and 3rd most popular physiological signals in Table 1. Blood pressure, the most popular physiological signal, was not attacked alone; it was considered only once in [85], together with EEG, ECG, etc. The

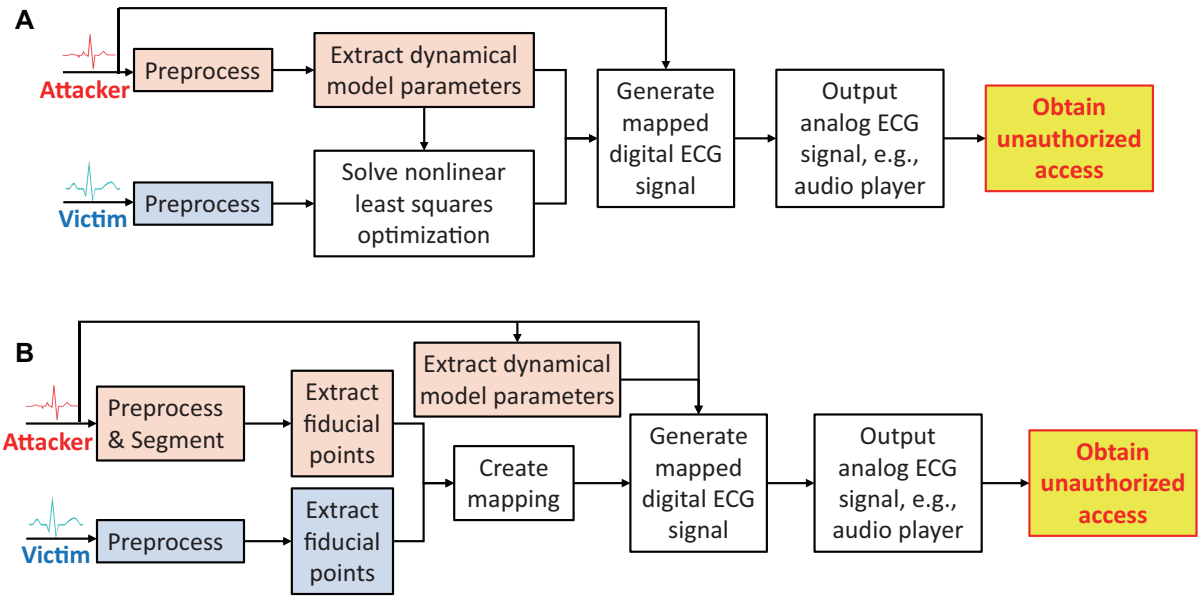

Figure 16 (A) Offline ECG biometrics presentation attack [88]; (B) online ECG biometrics presentation attack [95].

reason may be that blood pressure consists of only two numbers (systolic and diastolic pressures) measured infrequently, so it is not easy and interesting to attack. The same reasoning may also apply to respiration and heart rate. Some other physiological signals, e.g., ECoG and EMG, are frequently used in human-machine interactions, and also may be complex and important enough to attract adversarial attacks.

Finally, examining the current few studies on adversarial attacks of time series [75–77], we found that all of them did not take the causality of the time series into consideration, i.e., their approaches utilized the entire test trial in computing the adversarial perturbation, so they can only be used offline. Additionally, they used generic classifiers for all time series, whereas in physiological computing, particularly BCIs [12], each paradigm has its own best feature extraction and classification/regression approach, according to the neurological basis of the corresponding paradigm. So, these generic time series attack approaches may not be used directly in physiological computing.

Defense against adversarial attacks

There are different adversarial defense strategies [22, 96].

(1) Data modification, which modifies the training set in the training stage or the input data in the test stage, through adversarial training [20], gradient hiding [97], transferability blocking [98], data compression [99], data randomization [100], etc.

(2) Model modification, which modifies the target model directly to increase its robustness. This can be achieved through regularization [74], defensive distillation [101], feature squeezing [102], using a deep contractive network [103] or a mask layer [104], etc.

(3) Auxiliary tools, which may be additional auxiliary machine learning models to robustify the primary model, e.g., adversarial detection models [105], or defense generative adversarial nets (defense-GAN) [106], high-level representation guided denoiser [107], etc.

As researchers just started to investigate adversarial attacks in physiological computing, there were even fewer studies on defense strategies against them. A summary of them is shown in Table 5 [36, 76, 108–111].

Adversarial training

Adversarial training, which trains a robust machine learning model on normal plus adversarial examples, may be the most popular data modification based adversarial defense approach.

Hussein *et al.* [108] proposed an approach to augment deep learning models with adversarial training for

Table 5 Summary of existing adversarial defense studies in physiological computing

Ref.	Application	Data modification	Model modification	Adversarial detection
[108]	BCI	✓		
[109]	BCI		✓	
[110]	Health informatics			✓
[111]	Health informatics			✓
[76]	Health informatics	✓		
[36]	Biometrics			✓

robust prediction of epilepsy seizures. Though their goal was to overcome some challenges in EEG-based seizure classification, e.g., individual differences and shortage of pre-ictal labeled data, their approach can also be used to defend against adversarial attacks.

They first constructed a deep learning classifier from available limited amount of labeled EEG data, and then performed white-box attacks to the classifier to obtain adversarial examples, which were next combined with the original labeled data to retrain the deep learning classifier. Experiments on two public seizure datasets demonstrated that adversarial training increased both the classification accuracy and the classifier robustness.

More recently, Karim *et al.* [76] performed adversarial training for fully convolutional network classifiers, and tested the performance on 42 time series datasets, including three ECG datasets on heart conditions or Myocardial infarction classification, and two EOG datasets on Japanese Katakana stroke classification. They showed that even very simple adversarial training can improve the robustness of fully convolutional network classifiers to black-box and white-box non-targeted evasion attacks.

Although adversarial training may be the most effective approach for enhancing the robustness of a model, it could lead to undesirable accuracy degradation on the benign examples [112]. Additionally, it increases the computational cost 3–30 times [113].

Model modification

Regularization based model modification to defend against adversarial attacks usually considers the model security (robustness) in the optimization objective function.

Sadeghi *et al.* [109] proposed an analytical framework for tuning the classifier parameters, to ensure simultaneously its accuracy and security. The optimal classifier parameters were determined by solving an optimization problem, which takes into account both the test accuracy and the robustness against adversarial attacks. For k-nearest neighbor (kNN) classifiers, the two parameters to be optimized are the number of neighbors and the distance metric type. Experiments on EEG-based eye state (open or close) recognition verified that it is possible to achieve both high classification accuracy and high robustness against black-box targeted evasion attacks.

Model modification approaches are usually heuristic and empirical, without theoretical guarantees. They may be vulnerable to model-agnostic block-box attacks [69].

Adversarial detection

Adversarial detection uses a separate module to detect if there is adversarial attack, and takes actions accordingly. The simplest is to discard adversarial examples directly.

Cai and Venkatasubramanian [111] proposed an approach to detect signal injection-based morphological alterations (evasion attack) of ECGs. Because multiple physiological signals based on the same underlying physiological process (e.g., cardiac process) are inherently related to each other, any adversarial alteration of one of the signals will lead to inconsistency in the other signal(s) in the group. Since both ECG and arterial blood pressure measurements are representations of the cardiac process, the latter can be used to detect morphological alterations in ECGs. They demonstrated over 90% accuracy in detecting even subtle ECG morphological alterations for both healthy subjects and patients. A similar idea [110] was also used to

detect temporal alternations of ECGs, by making use of their correlations with arterial blood pressure and respiration measurements.

Karimian *et al.* [36] proposed two strategies to protect ECG biometric authentication systems from spoofing, by evaluating if ECG signal characteristics match the corresponding heart rate variability or PPG features (pulse transit time and pulse arrival time). The idea is actually similar to Cai and Venkatasubramanian's [111]. If there is a mismatch, then the system considers the input to be fake, and rejects it.

Adversarial detection heavily relies on the difference between adversarial examples and benign examples. However, adversarial examples can fool not only the classifier but also the detector, so adversarial detection may be ineffective against adaptive attacks [114].

Discussion

Although there have been many adversarial attack defense approaches [23], no one can withstand all existing attacks, not to mention new attacks that will for sure be discovered in the future. For example, Miller *et al.*'s experiments [23] showed that an adversarially trained robust deep neural network can accommodate small perturbations, at the cost of significant classification accuracy loss (about 10%) for benign inputs. Furthermore, as the attack strength increased, this robust classifier gradually became ineffective.

As proposed in [23], a promising new adversarial defense direction may be to combine robust classification with detection: robust classification forces the adversarial perturbations to be large for successful attacks, but this also makes the attacks more detectable. So, using robust classification and adversarial detection together may outperform each one alone.

Another idea is to use multi-modal inputs in physiological computing. Multi-modal signals are frequently used to increase the accuracy of physiological computing, e.g., using both EEG and EOG increased the emotion classification accuracy significantly [115]. They can also be used to increase the robustness to adversarial attacks, as different physiological signals generally require different perturbations, which raises the difficulty of attacking. However, this may also increase the complexity and cost of the resulting physiological computing system. Thus, there should be a careful trade-off between accuracy/robustness and complexity/cost.

Conclusions and future research

Physiological computing includes, or significantly overlaps with, BCIs, affective computing, adaptive automation, health informatics, and physiological signal based biometrics. It increases the communication bandwidth from the user to the computer, but is also subject to adversarial attacks. This study has given a comprehensive review on adversarial attacks and their defense strategies in physiological computing, hopefully will bring more attention to the security of physiological computing systems.

Promising future research directions in this area include the following.

(1) Transfer learning has been extensively used in physiological computing [12], to alleviate the training data shortage problem by leveraging data from other subjects [116] or tasks [117], or to warm-start the training of a (deep) learning algorithm by borrowing parameters or knowledge from an existing algorithm [86], as shown in Figure 17 [118]. However, transfer learning is particularly susceptive to poisoning attacks

[81,86]. It is very important to develop strategies to check the integrity of data and models before using them in transfer learning.

(2) Adversarial attacks to other components in the machine learning pipeline (an example on BCI is shown in Figure 18), which includes signal processing, feature engineering, classification/regression, and the corresponding defense strategies. So far all adversarial attack approaches in physiological computing considered the classification or regression model only, but not other components, e.g., signal processing and feature engineering. It has been shown that feature selection is also subjective to data poisoning attacks [72], and adversarial feature selection can be used to defend against evasion attacks [119].

(3) Additional types of attacks in physiological computing [96, 120–123], and the corresponding defense strategies, as shown in Figure 19. For example, Paoletti *et al.* [94] performed parameter tampering attacks on Boston Scientific implantable cardioverter defibrillators, which use a discrimination tree to detect tachycardia episodes and then initiate the appropriate therapy. They slightly modified the parameters of the discrimina-

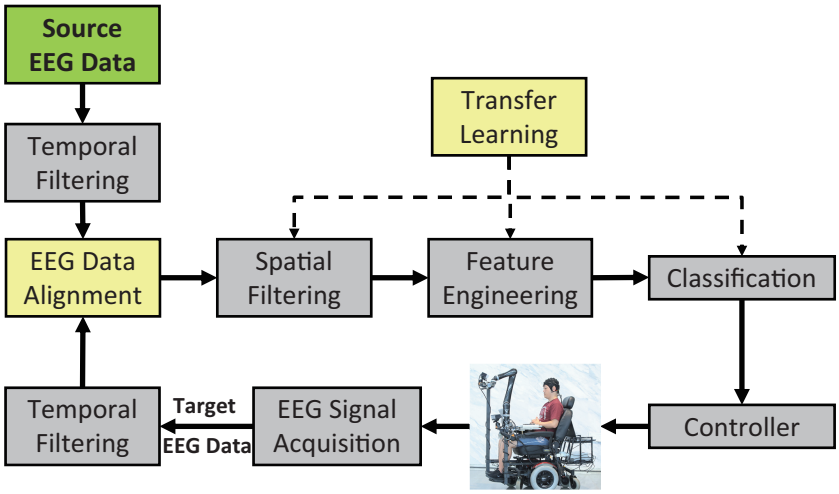

Figure 17 A transfer learning pipeline in motor imagery based BCIs [118].

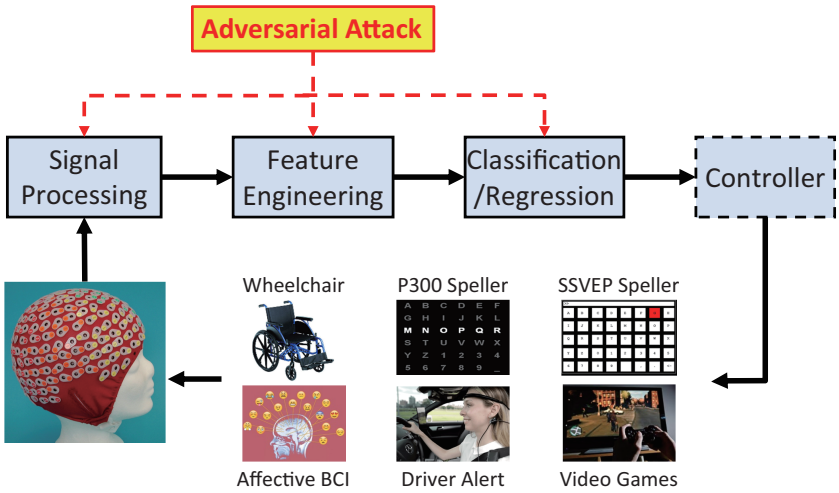

Figure 18 Adversarial attacks to the BCI machine learning pipeline.

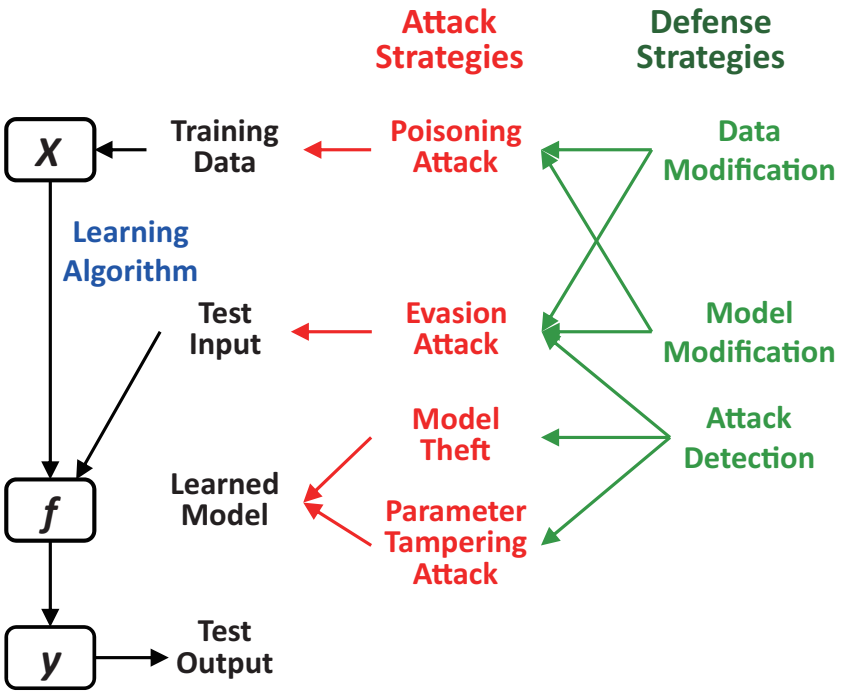

Figure 19 Additional types of attacks in physiological computing.

tion tree to achieve both attack effectiveness and stealthiness. These attacks are also very dangerous in physiological computing, and hence deserve adequate attention.

(4) Adversarial attacks to affective computing and adaptive automation applications, which have not been studied yet, but are also possible and dangerous. Many existing attack approaches in BCIs, health informatics and biometrics can be extended to them, either directly or with slight modifications. However, there could also be unique attack approaches specific to these areas. For example, emotions are frequently represented as continuous numbers in the 3D space of valence, arousal and dominance in affective computing [50], and hence adversarial attacks to regression models in affective computing should be paid enough attention to.

(5) Real-world demonstration of adversarial attacks and defenses. As mentioned in the Discussion, current research on adversarial attacks of time series did not consider their causality, so the attacks may not be used in the most meaningful online applications. Adversarial attacks to BCIs have advanced rapidly in the past few years, and the attacks have become very easy to perform in theory. However, real-world experiments are still needed to demonstrate their practicableness, and more importantly, the necessity, feasibility and benefits of adversarial defenses.

(6) Privacy of physiological computing systems. Adversarial attacked discussed in this review focused on manipulating the system outputs; however, privacy is another very important concern in physiological computing. For example, personal account, personal preferences, physical state and commercial models are privacy information that could be stolen from BCIs [124]. Defending against these privacy attacks is also crucial for wide-spread applications of physiological computing systems.

Finally, we need to emphasize that the goal of adversarial attack research in physiological computing should be discovering its vulnerabilities, and then finding solutions to make it more secure, instead of merely causing damages to it.

Funding

This work was supported by the Open Research Projects of Zhejiang Lab (2021KE0AB04), the Technology Innovation Project of Hubei Province of China (2019AEA171), the National Social Science Foundation of China (19ZDA104 and 20AZD089), and the Independent Innovation Research Fund of Huazhong University of Science and Technology (2020WKZDJC004).

Author contributions

D. Wu prepared the manuscript. All authors edited and proofread the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1 Fairclough SH. Fundamentals of physiological computing. *Interacting Comput* 2009; **21**: 133–145.
2 Minsky M. *The Society of Mind*. New York: Simon and Schuster, 1988
3 Jacucci G, Fairclough S, Solovey ET. Physiological computing. *Computer* 2015; **48**: 12–16.
4 Han X, Hu Y, Foschini L, *et al*. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nat Med* 2020; **26**: 360–363.
5 Lance BJ, Kerick SE, Ries AJ, *et al*. Brain-computer interface technologies in the coming decades. *Proc IEEE* 2012; **100**: 1585–1599.
6 Daly JJ, Wolpaw JR. Brain-computer interfaces in neurological rehabilitation. *Lancet Neurol* 2008; **7**: 1032–1043.
7 Huang H, Xie Q, Pan J, *et al*. An EEG-based brain computer interface for emotion recognition and its application in patients with disorder of consciousness. *IEEE Trans Affective Comput* 2021; **12**: 832–842.
8 Shanechi MM. Brain-machine interfaces from motor to mood. *Nat Neurosci* 2019; **22**: 1554–1564.
9 Chen X, Wang Y, Nakanishi M, *et al*. High-speed spelling with a noninvasive brain-computer interface. *Proc Natl Acad Sci USA* 2015; **112**: E6058–E6067.
10 Wolpaw JR, Birbaumer N, McFarland DJ, *et al*. Brain-computer interfaces for communication and control. *Clin Neurophysiol* 2002; **113**: 767–791.
11 Peng R, Jiang J, Kuang G, et al. EEG-based automatic Epilepsy detection: Review and outlook (in Chinese). *Acta Automatica Sinica*, 2022; **48**: 335–350.
12 Wu D, Xu Y, Lu BL. Transfer learning for EEG-based brain-computer interfaces: A review of progress made since 2016. *IEEE Trans Cogn Dev Syst* 2022; **14**: 4–19.
13 Rim B, Sung NJ, Min S, *et al*. Deep learning in physiological signal data: A survey. *Sensors* 2020; **20**: 969.
14 Lawhern VJ, Solon AJ, Waytowich NR, *et al*. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *J Neural Eng* 2018; **15**: 056013.
15 Schirrmeister RT, Springenberg JT, Fiederer LDJ, *et al*. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum Brain Mapp* 2017; **38**: 5391–5420.
16 Kostas D, Rudzicz F. Thinker invariance: Enabling deep neural networks for BCI across more people. *J Neural Eng* 2020; **17**: 056008.
17 Asif U, Roy S, Tang J, *et al*. SeizureNet: Multi-spectral deep feature learning for seizure type classification. In: *Proceedings of Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*, 2020. 77–87.
18 Goodfellow SD, Goodwin A, Greer R, *et al*. Towards understanding ECG rhythm classification using convolutional neural networks and attention mappings. In: *Proceedings of the 3rd Machine Learning for Healthcare Conference*, Stanford, 2018. 83–101.
19 Hwang B, You J, Vaessen T, *et al*. Deep ECGNet: An optimal deep learning framework for monitoring mental stress using ultra short-term ECG signals. *TeleMed e-Health* 2018; **24**: 753–772.

20 Szegedy C, Zaremba W, Sutskever I, *et al*. Intriguing properties of neural networks. In: *Proceedings of International Conference on Learning Representations*, Banff, 2014.

21 Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: *Proceedings of International Conference on Learning Representations*, San Diego, 2015.

22 Qiu S, Liu Q, Zhou S, *et al*. Review of artificial intelligence adversarial attack and defense technologies. *Appl Sci* 2019; **9**: 909.

23 Miller DJ, Xiang Z, Kesidis G. Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks. *Proc IEEE* 2020; **108**: 402–433.

24 Sharif M, Bhagavatula S, Bauer L, *et al*. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, Vienna, 2016. 1528–1540.

25 Brown TB, Mané D, Roy A, *et al*. Adversarial patch. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 2017.

26 Chen Q, Ma X, Zhu Z, *et al*. Evolutionary multi-tasking single-objective optimization based on cooperative co-evolutionary memetic algorithm. In: *Proceedings of the 13th International Conference on Computational Intelligence and Security*, 2017. 197–201.

27 Athalye A, Engstrom L, Ilyas A, *et al*. Synthesizing robust adversarial examples. In: *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, 2018. 284–293.

28 Evtimov I, Eykholt K, Fernandes E, *et al*. Robust physical-world attacks on deep learning visual classiffication. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018. 1625–1634.

29 Finlayson SG, Chung HW, Kohane IS, *et al*. Adversarial attacks against medical deep learning systems. ArXiv: 1804.05296.

30 Finlayson SG, Bowers JD, Ito J, *et al*. Adversarial attacks on medical machine learning. *Science* 2019; **363**: 1287–1289.

31 Rahman A, Hossain MS, Alrajeh NA, *et al*. Adversarial examples-security threats to COVID-19 deep learning systems in medical IoT devices. *IEEE Internet Things J* 2021; **8**: 9603–9610.

32 Ma X, Niu Y, Gu L, *et al*. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition* 2021; **110**: 107332.

33 Kaissis GA, Makowski MR, Rückert D, *et al*. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell* 2020; **2**: 305–311.

34 Zhang X, Wu D, Ding L, *et al*. Tiny noise, big mistakes: Adversarial perturbations induce errors in brain-computer interface spellers. *Natl Sci Rev* 2021; **8**: nwaa233.

35 Karimian N. How to attack PPG biometric using adversarial machine learning. In: *Proceedings of Autonomous Systems: Sensors, Processing, and Security for Vehicles and Infrastructure*, 2019. 11009: 1100909

36 Karimian N, Woodard D, Forte D. ECG biometric: Spoofing and countermeasures. *IEEE Trans Biom Behav Identity Sci* 2020; **2**: 257–270.

37 Bernal SL, Celdran AH, Maimó LF, *et al*. Cyberattacks on miniature brain implants to disrupt spontaneous neural signaling. *IEEE Access* 2020; **8**: 152204.

38 Pfurtscheller G, Neuper C. Motor imagery and direct brain-computer communication. *Proc IEEE* 2001; **89**: 1123–1134.

39 Handy TC. *Event-Related Potentials: A Methods Handbook*. Boston: The MIT Press, 2005

40 Lees S, Dayan N, Cecotti H, et al. A review of rapid serial visual presentation-based brain-computer interfaces. *J Neural Eng* 2018; **15**: 021001.

41 Sutton S, Braren M, Zubin J, *et al*. Evoked-potential correlates of stimulus uncertainty. *Science* 1965; **150**: 1187–1188.

42 Friman O, Volosyak I, Graser A. Multiple channel detection of steady-state visual evoked potentials for brain-computer

interfaces. *IEEE Trans Biomed Eng* 2007; **54**: 742–750.

43 Beverina F, Palmas G, Silvoni S, *et al*. User adaptive BCIs: SSVEP and P300 based interfaces. *Psychnology J*, 2003; **1**: 331–354

44 Sellers EW, Donchin E. A P300-based brain-computer interface: Initial tests by ALS patients. *Clin Neurophysiol* 2006; **117**: 538–548.

45 Geller EB. Responsive neurostimulation: Review of clinical trials and insights into focal epilepsy. *Epilepsy Behav* 2018; **88**: 11–20.

46 Gummadavelli A, Zaveri HP, Spencer DD, *et al*. Expanding brain-computer interfaces for controlling epilepsy networks: Novel thalamic responsive neurostimulation in refractory epilepsy. *Front Neurosci* 2018; **12**: 474.

47 Picard R. *Affective Computing*. Cambridge: The MIT Press, 1997.

48 Ekman P, Friesen WV. Constants across cultures in the face and emotion. *J Personality Soc Psychol* 1971; **17**: 124–129.

49 Russell JA. A circumplex model of affect. *J Personality Soc Psychol* 1980; **39**: 1161–1178.

50 Mehrabian A. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Cambridge: Oelgeschlager, Gunn & Hain, 1980.

51 Quan X, Zeng Z, Jiang J, *et al*. Physiological signals based affective computing: A systematic review (in Chinese). *Acta Automatica Sinica*, 2021; **47**: 1769–1784.

52 Chittaro L, Sioni R. Affective computing vs. affective placebo: Study of a biofeedback-controlled game for relaxation training. *Int J Hum-Comput Studies* 2014; **72**: 663–673.

53 Aranha RV, Correa CG, Nunes FLS. Adapting software with affective computing: A systematic review. *IEEE Trans Affective Comput* 2021; **12**: 883–899.

54 Boeke DK, Miller ME, Rusnock CF, *et al*. Exploring individualized objective workload prediction with feedback for adaptive automation. In: *Proceedings of Industrial and Systems Engineering Research Conference*, Nashville, 2015. 1437–1446.

55 Aricó P, Borghini G, Di Flumeri G, *et al*. Adaptive automation triggered by EEG-based mental workload index: A passive brain-computer interface application in realistic air traffic control environment. *Front Hum Neurosci* 2016; **10**: 539.

56 de Greef T, Lafeber H, van Oostendorp H, *et al*. Eye movement as indicators of mental workload to trigger adaptive automation. In: *Proceedings of International Conference on Foundations of Augmented Cognition*, San Diego, 2009. 219–228.

57 Park J, Zahabi M. Cognitive workload assessment of prosthetic devices: A review of literature and meta-analysis. *IEEE Trans Hum-Mach Syst* 2022; **52**: 181–195.

58 Coiera E. *Guide to Health Informatics*. Boca Raton: CRC Press, 2015

59 Mishra T, Wang M, Metwally AA, *et al*. Pre-symptomatic detection of COVID-19 from smartwatch data. *Nat Biomed Eng* 2020; **4**: 1208–1220.

60 Charlton PH, Kyriacou PA, Mant J, *et al*. Wearable photoplethysmography for cardiovascular monitoring. *Proc IEEE* 2022; **110**: 355–381.

61 Guo YT, Cui Y, Zhao C, *et al*. Machine-learning fusion approach for the prediction of atrial fibrillation onset using photoplethysmographic-based smart device. *Eur Heart J* 2021; **42**: ehab724.3058.

62 Singh YN, Singh SK, Ray AK. Bioelectrical signals as emerging biometrics: Issues and challenges. *ISRN Signal Processing* 2012; **2012**: 1–13.

63 Thomas KP, Vinod AP. Toward EEG-Based biometric systems: The great potential of brain-wave-based biometrics. *IEEE Syst Man Cybern Mag* 2017; **3**: 6–15.

64 Agrafioti F, Gao J, Hatzinakos D, *et al*. Heart Biometrics: Theory, Methods and Applications. In: *Biometrics*. London: InTechOpe, 2011. 199–216.

65 Yadav U, Abbas SN, Hatzinakos D. Evaluation of PPG biometrics for authentication in different states. In: *Proceedings*

of International Conference on Biometrics, Queensland, 2018. 277–282.

66 Bianco S, Napoletano P. Biometric recognition using multimodal physiological signals. *IEEE Access* 2019; **7**: 83581–83588.

67 Zhang X, Wu D. On the vulnerability of CNN classifiers in EEG-based BCIs. *IEEE Trans Neural Syst Rehabil Eng* 2019; **27**: 814–825.

68 Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016. 2574–2582.

69 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: *Proceedings of IEEE Symposium on Security and Privacy*, San Jose, 2017. 39–57.

70 Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. In: *Proceedings of International Conference on Learning Representations*, Toulon, 2017.

71 Papernot N, McDaniel P, Goodfellow I, *et al*. Practical black-box attacks against machine learning. In: *Proceedings of Asia Conference on Computer and Communications Security*, Abu Dhabi, 2017. 506–519.

72 Xiao H, Biggio B, Brown G, *et al*. Is feature selection secure against training data poisoning? In: *Proceedings of the 32nd International Conference on Machine Learning*, Lille, 2015. 1689–1698.

73 Mei S, Zhu X. Using machine teaching to identify optimal training-set attacks on machine learners. In: *Proceedings of AAAI Conference on Artificial Intelligence*, 2015. 29: 2871–2877.

74 Biggio B, Nelson B, Laskov P. Support vector machines under adversarial label noise. In: *Proceedings of Asian Conference on Machine Learning*, Taipei, 2011. 97–112.

75 Fawaz HI, Forestier G, Weber J, *et al*. Adversarial attacks on deep neural networks for time series classification. In: *Proceedings of International Joint Conference on Neural Networks*, Budapest, 2019. 1–8.

76 Karim F, Majumdar S, Darabi H. Adversarial attacks on time series. *IEEE Trans Pattern Anal Mach Intell* 2021; **43**: 3309–3320.

77 Harford S, Karim F, Darabi H. Generating adversarial samples on multivariate time series using variational autoencoders. *IEEE CAA J Autom Sin* 2021; **8**: 1523–1538.

78 Cheng P, Roedig U. Personal voice assistant security and privacy—A survey. *Proc IEEE* 2022; **110**: 476–507.

79 Jiang X, Zhang X, Wu D. Active learning for black-box adversarial attacks in EEG-based brain-computer interfaces. In: *Proceedings of IEEE Symposium Series on Computational Intelligence*, Xiamen, 2019.

80 Liu Z, Meng L, Zhang X, *et al*. Universal adversarial perturbations for CNN classifiers in EEG-based BCIs. *J Neural Eng* 2021; **18**: 0460a4.

81 Meng L, Huang J, Zeng Z, *et al*. EEG-based brain-computer interfaces are vulnerable to backdoor attacks. *Engineering*, 2022, doi: 10.21203/rs.3.rs-108085/v1.

82 Bian R, Meng LB, Wu DR. SSVEP-based brain-computer interfaces are vulnerable to square wave attacks. *Sci China Inf Sci* 2022; **65**: 140406.

83 Meng L, Lin C-T, Jung T-P, *et al*. White-box target attack for EEG-based BCI regression problems. In: *Proceedings of International Conference on Neural Information Processing*, Sydney, 2019.

84 Aminifar A. Universal adversarial perturbations in epileptic seizure detection. In: *Proceedings of International Joint Conference on Neural Networks*, 2020. 1–6.

85 Newaz A, Haque NI, Sikder AK, *et al*. Adversarial attacks to machine learning-based smart healthcare systems. ArXiv: 2010.03671.

86 Wang S, Nepal S, Rudolph C, *et al*. Backdoor attacks against transfer learning with pre-trained deep learning models. *IEEE Trans Serv Comput* 2022; **15**: 1526–1539.

87 Maiorana E, Hine GE, Rocca DL, et al. On the vulnerability of an EEG-based biometric system to hill-climbing attacks algorithms' comparison and possible countermeasures. In: *Proceedings of IEEE 6th International Conference on Biometrics: Theory, Applications and Systems*, 2013. 1–6.

88 Eberz S, Paoletti N, Roeschlin M, *et al*. Broken hearted: How to attack ECG biometrics. In: *Proceedings of Network and Distributed System Security Symposium*. San Diego: Internet Society, 2017.

89 Wu D, Lawhern VJ, Gordon S, *et al*. Driver drowsiness estimation from EEG signals using online weighted adaptation regularization for regression (OwARR). *IEEE Trans Fuzzy Syst* 2017; **25**: 1522–1535.

90 Ienca M, Haselager P, Emanuel EJ. Brain leaks and consumer neurotechnology. *Nat Biotechnol* 2018; **36**: 805–810.

91 Jarchum I. The ethics of neurotechnology. *Nat Biotechnol* 2019; **37**: 993–996.

92 Binnendijk A, Marler T, Bartels EM. *Brain-Computer Interfaces: U.S. Military Applications and Implications, An Initial Assessment*. Santa Monica: RAND Corporation, 2020

93 Sundararajan K. Privacy and security issues in brain computer interfaces. Dissertation for Master's Degree. Auckland: Auckland University of Technology, 2017.

94 Paoletti N, Jiang Z, Islam MA, *et al*. Synthesizing stealthy reprogramming attacks on cardiac devices. In: *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, 2019. 13–22.

95 Karimian N, Woodard DL, Forte D. On the vulnerability of ECG verification to online presentation attacks. In: *Proceedings of IEEE International Joint Conference on Biometrics*, Denver, 2017. 143–151.

96 Bernal SL, Celdrán AH, Pérez GM, *et al*. Security in brain-computer interfaces. *ACM Comput Surv* 2022; **54**: 1–35.

97 Tramèr F, Kurakin A, Papernot N, *et al*. Ensemble adversarial training: Attacks and defenses. ArXiv: 1705.07204.

98 Hosseini H, Chen Y, Kannan S, *et al*. Blocking transferability of adversarial examples in black-box learning systems. ArXiv: 1703.04318.

99 Das N, Shanbhogue M, Chen S-T, *et al*. Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression. ArXiv: 1705.02900.

100 Xie C, Wang J, Zhang Z, *et al*. Adversarial examples for semantic segmentation and object detection. In: *Proceedings of IEEE International Conference on Computer Vision*, Venice, 2017. 1369–1378.

101 Papernot N, McDaniel P, Wu X, *et al*. Distillation as a defense to adversarial perturbations against deep neural networks. In: *Proceedings of IEEE Symposium on Security and Privacy*, San Jose, 2016. 582–597.

102 Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks. ArXiv: 1704.01155.

103 Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. ArXiv: 1412.5068.

104 Gao J, Wang B, Lin Z, *et al.* DeepCloak: Masking deep neural network models for robustness against adversarial samples. ArXiv: 1702.06763.

105 Qayyum A, Qadir J, Bilal M, *et al*. Secure and robust machine learning for healthcare: A survey. *IEEE Rev Biomed Eng* 2021; **14**: 156–180.

106 Samangouei P, Kabkab M, Chellappa R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. ArXiv: 1805.06605.

107 Liao F, Liang M, Dong Y, *et al*. Defense against adversarial attacks using high-level representation guided denoiser. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018. 1778–1787.

108 Hussein A, Djandji M, Mahmoud RA, *et al*. Augmenting DL with adversarial training for robust prediction of epilepsy seizures. *ACM Trans Comput Healthcare* 2020; **1**: 1–18.

109 Sadeghi K, Banerjee A, Gupta SK. An analytical framework for security-tuning of artificial intelligence applications under attack. In: *Proceedings of IEEE International Conference On Artificial Intelligence Testing*, San Francisco, 2019. 111–118.

110 Cai H, Venkatasubramanian KK. Detecting malicious temporal alterations of ECG signals in body sensor networks. In: *Proceedings of International Conference on Network and System Security*, New York, 2015. 531–539.

111 Cai H, Venkatasubramanian KK. Detecting signal injection attack-based morphological alterations of ECG measurements. In: *Proceedings of International Conference on Distributed Computing in Sensor Systems*, Washington, 2016. 127–135.

112 Rade R, Moosavi-Dezfooli S-M. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In:

Proceedings of International Conference on Learning Representations, 2022.

113 Shafahi A, Najibi M, Ghiasi A, et al. Adversarial training for free! In: *Proceedings of Advances in Neural Information Processing Systems*, Vancouver, 2019.

114 Carlini N, Wagner DA. Adversarial examples are not easily detected: Bypassing ten detection methods. In: *Proceedings of Workshop on Artificial Intelligence and Security*, Dallas, 2017.

115 Zheng WL, Liu W, Lu Y, *et al*. EmotionMeter: A multimodal framework for recognizing human emotions. *IEEE Trans Cybern* 2019; **49**: 1110–1122.

116 He H, Wu D. Transfer learning for brain-computer interfaces: A Euclidean space data alignment approach. *IEEE Trans Biomed Eng* 2020; **67**: 399–410.

117 Wu D, Lawhern VJ, Hairston WD, *et al*. Switching EEG headsets made easy: Reducing offline calibration effort using active weighted adaptation regularization. *IEEE Trans Neural Syst Rehabil Eng* 2016; **24**: 1125–1137.

118 Wu D, Jiang X, Peng R. Transfer learning for motor imagery based brain-computer interfaces: A tutorial. *Neural Networks* 2022; **153**: 235–253.

119 Zhang F, Chan PPK, Biggio B, *et al*. Adversarial feature selection against evasion attacks. *IEEE Trans Cybern* 2016; **46**: 766–777.

120 Denning T, Matsuoka Y, Kohno T. Neurosecurity: Security and privacy for neural devices. *FOC* 2009; **27**: E7.

121 Rushanan M, Rubin AD, Kune DF, *et al*. SoK: Security and privacy in implantable medical devices and body area networks. In: *Proceedings of IEEE Symposium on Security and Privacy*, 2014. 524–539.

122 Camara C, Peris-Lopez P, Tapiador JE. Security and privacy issues in implantable medical devices: A comprehensive survey. *J BioMed Inf* 2015; **55**: 272–289.

123 Pycroft L, Boccard SG, Owen SLF, *et al*. Brainjacking: Implant security issues in invasive neuromodulation. *World Neurosurg* 2016; **92**: 454–462.

124 Xia K, Duch W, Sun Y, *et al*. Privacy-preserving brain-computer interfaces: A systematic review. *IEEE Trans Comput Soc Syst* 2022; doi: 10.1109/TCSS.2022.3184818.