Cancellable Deep Learning Framework for EEG Biometrics

Min Wang, *Member, IEEE*, Xuefei Yin, and Jiankun Hu, *Senior Member, IEEE*

***Abstract*— EEG-based biometric systems verify the identity of a user by comparing the probe to a reference EEG template of the claimed user enrolled in the system, or by classifying the probe against a user verification model stored in the system. These approaches are often referred to as template-based and model-based methods, respectively. Compared with template-based methods, model-based methods, especially those based on deep learning models, tend to provide enhanced performance and more flexible applications. However, there is no public research report on the security and cancellability issue for model-based approaches. This becomes a critical issue considering the growing popularity of deep learning in EEG biometric applications. In this study, we investigate the security issue of deep learning model-based EEG biometric systems, and demonstrate that model inversion attacks post a threat for such model-based systems. That is to say, an adversary can produce synthetic data based on the output and parameters of the user verification model to gain unauthorized access by the system. We propose a cancellable deep learning framework to defend against such attacks and protect system security. The framework utilizes a generative adversarial network to approximate a non-invertible transformation whose parameters can be changed to produce different data distributions. A user verification model is then trained using output generated from the generator model, while information about the transformation is discarded. The proposed framework is able to revoke compromised models to defend against hill climbing attacks and model inversion attacks. Evaluation results show that the proposed method, while being cancellable, achieves better verification performance than the template-based methods and state-of-the-art non-cancellable deep learning methods.**

***Index Terms*— Cancellable biometrics, deep learning, EEG biometrics, user verification, neural networks, biometric security.**

I. Introduction

BRAIN biometrics based on electroencephalography (EEG) have garnered increasing attention due to the potential security benefits, e.g., enhanced robustness against circumvention and support for intrinsic liveness detection [1]. EEG signals generated by cerebral activity are internal traits that are hidden from public access, and the acquisition of EEG biometrics requires the conscious and cooperative engagement of the user [2]. Therefore, it is highly unlikely to capture EEG signals of a target user covertly or remotely without the user's awareness, making EEG biometrics less susceptible to sensor spoofing attacks. Moreover, as many features of EEG signals are non-volitional [3], meaning they are beyond the control or conscious apprehension of the user, it protects the biometric identifiers from deliberate disclosure. In addition, EEG, as an indicator of brain activity and liveness [3], naturally provides liveness detection capabilities and reduces the possibility of presentation attacks using spoofing artifacts or lifeless body parts. These peculiarities offer potentials for more flexible and secure biometric systems [4]. However, more efforts are expected to bring these advantages to fruition.

EEG biometric systems generally comprise three main components: a signal acquisition module, a feature extraction module, and a template comparison or classification module [3], [4], [5]. Recently, deep learning-based approaches are fast-growing, where neural networks of different types, architectures, and schemes are developed to automatically learn high-level representations of EEG from the data for biometric identification and verification [1], [6]. We can categorize them into template comparison-based approaches and model-based approaches. Template-based methods store a biometric template for each user during the enrollment phase and compare the probe against the template stored for the claimed identity during the verification phase to decide whether to accept or reject the request. A model-based approach, instead of storing a template, trains and saves a classification model in the system during the enrollment phase and then uses the model for predictions during verification. Fig. 1 depicts the differences between the two types of EEG biometric systems. The question is: is it secure to directly store biometric templates or classification models in EEG biometric systems?

The answer is no for template-based approaches. Studies have shown that raw EEG signals and EEG templates (features) used for biometric applications reveal personal characteristics of users, including age and gender, as well as sensitive information related to drug intake, neurological disorders, and cognitive and mental states [7]. Therefore, when the templates stored in the system are stolen or obtained illegally by attackers, there will be a serious risk of user privacy leakage. To address this issue, various privacy-preserving mechanisms were proposed to protect EEG templates, including cryptographic schemes based on hash functions [8] and fuzzy commitment [9]. However, these approaches are not cancellable, meaning the system cannot cancel and revoke

Manuscript received 20 November 2023; accepted 2 February 2024. Date of publication 23 February 2024; date of current version 26 April 2024. This work was supported by the Australian Research Council under Grant DP200103207. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhen Lei. *(Corresponding author: Jiankun Hu.)*

Min Wang is with the School of Information Technology and Systems, University of Canberra, ACT 2617, Australia, and also with the School of Systems and Computing, University of New South Wales, Canberra, ACT 2612, Australia (e-mail: min.wang@canberra.edu.au).

Xuefei Yin is with the School of Information and Communication Technology, Griffith University, Gold Coast, QLD 4222, Australia (e-mail: x.yin@griffith.edu.au).

Jiankun Hu is with the School of Systems and Computing, University of New South Wales, Canberra, ACT 2612, Australia (e-mail: j.hu@adfa.edu.au).

Digital Object Identifier 10.1109/TIFS.2024.3369405

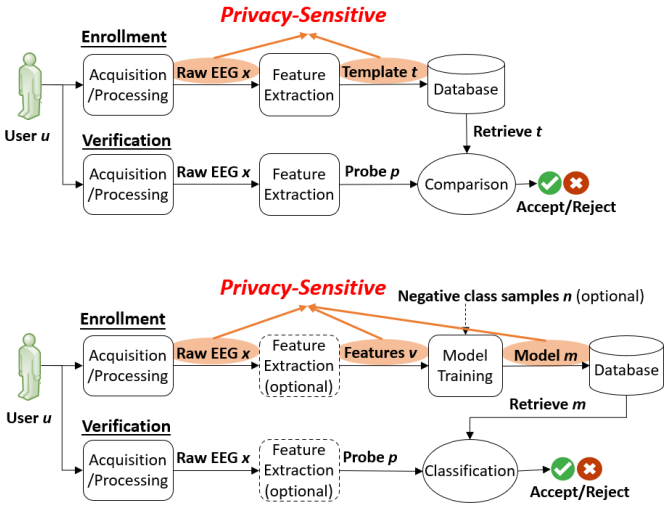


Fig. 1. Template-based systems (top) and model-based systems (bottom).

a compromised template in case of a security breach. This revocability issue was further addressed by recent studies [10], [11], which proposed cancellable template design for EEG biometrics by using non-invertible transformations. The transformation converts a raw biometric template into an 'encrypted' template to be stored in the system so that template comparison is performed in the transformed (encrypted) domain. More importantly, if the stored template is compromised in a security breach, the system is able to revoke the compromised template and issue a new one by changing the parameters (e.g., the key) of the transformation.

For model-based approaches, the answer is not confirmatory. Previous studies in EEG biometrics consider deep neural network-based verification model a secure procedure. For example, Bidgoly et al. [12] analogized deep learning model as a 'hashing' process that hides the user's private information and protect user's raw EEG signals. However, this may not be true since recent studies in machine learning suggest that a deep learning model can leak information about its training data through its output and parameters [13]. An adversary can produce synthetic data based on the output and parameters of the user verification model and produce synthetic data to gain unauthorized access by the system. This type of attack is often referred to as model inversion attack [14], [15], [16]. Unfortunately, there is no public research report on the security issue for model-based EEG biometric systems. This becomes a critical issue considering the growing popularity of deep learning in EEG biometric applications.

In this study, we investigate the security issue of deep learning model-based EEG biometric systems. Specially, we launch the model inversion attack in a black-box setting [13], [17], which assumes the attacker can submit queries to the model and get the corresponding confidence scores but do not have any available data from the users. This assumption is applicable because the authentication system is usually deployed in the user device, which can be stolen and compromised by an attacker to gain access to the target model [18], [19]. Model inversion attacks under black-box settings and more aggressive white-box settings have been widely discussed in biometric systems based on face recognition [13], [14], [15]. Our experimental results demonstrate that model inversion attacks also post a threat for deep learning model-based EEG biometric systems, i.e., an adversary can launch such attacks to generate synthetic data to gain unauthorized access to the system. As a synthetic sample that allows false acceptance will compromise the whole system, it is insecure to directly train a user verification model and store the model in the authentication system without a protection mechanism. Therefore, in this paper, we propose a cancellable deep learning framework to protect EEG biometric systems and offer revocability capabilities to the classification model. This paper is a pioneering study on cancellable biometric design for deep learning-based approaches and provides insights for future research in this direction. Our contributions can be summarized as follows:

- Previous studies consider deep learning model-based approaches secure for EEG-based biometric systems. We launch model inversion attacks and demonstrate that model-based systems are vulnerable to such attacks; thus, additional security mechanisms are needed.
- We propose a novel concept of biometric cancellability: distribution cancellability, which alters the raw EEG data distribution through cancellable transformation before training a predictive model so that the model can be revoked. Based on the concept, a cancellable deep learning framework is proposed for EEG biometrics. The framework is able to revoke compromised models in the event of a security breach and issue new models to restore normal functionality for user authentication.
- Technically, the framework utilizes a GAN model to approximate a non-invertible transformation whose parameters can be changed to produce different distribution transforms. A classifier is then trained for user verification using data generated from the generator, while information about the transformation is discarded. Only the generator and classifier are saved in the system. When the stored model is compromised, we can change the transformation parameters to derive a new model for the user.

The remaining content of this paper is organized as follows: Section II reviews related works on EEG biometrics, relevant attacks, and the security developments; Section III presents the threat model and evaluation results of existing deep learning-based EEG biometric verification models under the attack; Section IV elaborates the proposed cancellable deep learning framework for EEG biometric authentication, followed by experiments and evaluation results in Section V. Section VI concludes the study and discusses future research directions.

II. Related Work

A. EEG Biometrics

Existing research on EEG biometrics mainly focuses on the optimization of signal acquisition protocols, feature extraction methods, and classification algorithms to improve biometric

performance. Popular signal acquisition protocols been investigated include the resting state [20], motor imagery tasks [21], event-related potential tasks [3], visual stimulation-based protocols, and user-defined tasks such as the pass-thoughts [22]. In terms of feature extraction, many EEG features are extracted from the time, frequency, phase and coherence domains, including parameters of autoregressive models [23], entropy measures [24], coefficients of Mel-frequency cepstrum, connectivity features [5], [25], and EEG spectral features derived from fast Fourier analysis [23], [26], wavelet packet decomposition [6] and other time-frequency analysis [4]. These features are combined into a biometric template for template comparison, where a distance or similarity measure (e.g. Manhattan distance [23]) is adopted to calculate a matching score that is compared with a threshold to decide whether to accept or reject the probe. Alternatively, classification models can be established to make predictions. In such cases, a training set (of raw EEG signals or features) is created during the enrollment phase and used to train a model for classification. Various algorithms have been proposed for classification based on linear discriminant analysis [6], support vector machines and neural networks [1], [27]. In particular, as data collection becomes more accessible, deep learning models are having an increasing impact on EEG biometrics due to their dramatic improvements in classification accuracy. Popular models include multilayer perceptron (MLP), convolutional neural networks (CNNs), long-short-term-memory (LSTM) models, graph neural networks [1], [6]. These models can capture high-level representations of EEG signals related to identity-bearing information, thus improving biometric performance. However, both template-based and model-based methods have security risks, and are potentially vulnerable to model inversion attacks and hill-climbing attacks [18], [28].

B. Model Inversion Attacks and Hill Climbing Attacks for Biometrics

In EEG biometrics, existing studies often consider deep neural network a secure procedure. For example, Bidgoly et al. [12] analogized deep learning model as a hashing process and considered that it is safe to directly store deep learning model in the verification system. However, this may not be true because the resulting models may be used by attacker to gain unauthorized access to the system.

Model inversion attacks aim to exploit the correlation between the input data and the model output to reconstruct sensitive features of the training data [16]. This is typically done by formulating an optimization problem to find the input values that maximize the likelihood under the target model. So far, effective model inversion attacks have been demonstrated on simple models such as linear regression, but remain challenging on deep neural networks due to the intractable and ill-conditioned nature of the underlying attack optimization problem [14]. For a deep neural network, the input to be recovered lies in a high-dimensional and continuous data space and directly optimizing over the high-dimensional space without constraints will fail to obtain true input. Currently, most of the studies on model inversion related to biometrics target shallow to medium-scale neural networks in the context of face recognition [16]. However, the performance of such attacks are not satisfactory because the recovered face images are often blurry, unrealistic or unrecognizable. Meanwhile, the reconstruction quality dramatically degrades with increasing complexity of the model architecture. To improve model inversion performance, additional constraints drawn from semantic information and auxiliary knowledge are applied in the optimization procedure [14], and generative models such as generative adversarial networks (GANs) and the variants are used for face image generation [15]. Although model inversion attacks have been discussed in face recognition [14], [15] and speaker recognition [19], there is so far no report on evaluation of model inversion attacks for EEG biometrics.

Hill climbing attack is a closely related but different concept from model inversion attack in biometrics. In a hill-climbing attack, the adversary exploit the confidence scores produced by the system with the goal of generating synthetic data that can allow a false acceptance [18]. It has been identified as a security threat to EEG biometric systems, especially template-based approaches [28]. To defend, cancellable template design based on non-invertible transformation has been proposed [10], [11].

C. Security Mechanisms

For template based EEG biometric systems, researchers have proposed security-enhancing algorithms based on hash functions, fuzzy commitment and error-correcting codes to protect the raw biometric templates. In these work, EEG features are extracted and then encrypted by hash functions (e.g., the fast Johnson-Lindenstrauss algorithm [8]) or hidden through cryptographic schemes (e.g., a fuzzy commitment construct [9]). Moreover, deep neural network were used to generate EEG templates and such feature extraction process are analogized as a hashing process that hides users' private information [12]. Another method [29] applies turbo codes and modulation constellations to generate codewords and binds EEG features with the codeword to derive a template. The binding operator reveal no information about its arguments, thus protecting EEG data. These methods, although privacy-preserving, are not cancellable, and thus cannot resist hill-climbing attacks and secondary attacks. With a synthetic sample obtained through hill-climbing attacks, the whole system is comprised. To defend, the system needs to be privacy-preserving as well as cancellable, which means: 1) the stored templates should not reveal any sensitive information about the raw biometrics; and 2) the system should be able to revoke compromised templates and issue new templates. Recent studies proposed cancellable template designs for EEG biometrics based on non-invertible transformations, e.g., the multivariate polynomial functions [10], [11]. Specially, the algorithms transform a biometric template (EEG features) into an 'encrypted' template that preserves EEG privacy. The transformation is non-invertible, thus recovering the raw biometric template from the transformed template is infeasible. Moreover, uncorrelated templates can be derived from the same biometric template by changing the parameter of the transformation. Therefore, the algorithm can revoke and replace a compromised template with a new one. However,

so far, there has been no related work providing a cancellability function for model-based EEG biometric systems.

III. Threat Model

This session introduces the threat model of model inversion attacks for biometric systems and the evaluation results of existing deep learning-based EEG biometric verification models under such attacks.

A. Threat Model

In EEG-based biometric authentication, private data of the user is used to train a deep learning model for user verification. The model inversion attack is a representative attack on machine learning models, which infers the training data of a model by accessing the model. Depending on the knowledge and capacity of the attacker, model inversion attacks can be launched in three settings, which are the white-box setting where all parameters of the model are accessible to the attacker, black-box setting where attacker can access soft inference output consisting of confidence scores [17], and label-only setting where only inference results in hard label forms are available. In terms of the threat model and experimental setup, model inversion attacks with black-box settings share similarities with the hill-climbing attacks [18] as both of them aim to obtain synthetic input of the verification model to gain false acceptance by exploiting the confidence scores of the model. We refer to the verification model subject to attacks as the target model, and focus on the black-box setting which is a reasonable and applicable condition for EEG biometrics that has been discussed in previous studies [18], [28].

1) Adversary's Goal: Let T denote the target verification model trained with a private dataset D, where $T : x \rightarrow \{s_0, s_1\}, s \in [0, 1]$, is the established mapping from an EEG input $x \in \mathbb{R}^d$ to a class, either 1 (user class) or 0 (non-user/impostor class). The goal of the model inversion attacks is to reconstruct representative data of the user class from the target model T.

2) Adversary's Power: Adversaries are aware of the purpose of the target model, as information on the task of the target model can be easily inferred from the application or classes of the model output [17], [30]. They can submit queries q to the model and access the corresponding confidence scores $s \leftarrow T(q)$. This setting is similar to the one used in the hill-climbing attack in biometrics [18].

B. Invasion Model

We implement the model inversion attack using a generative model-based approach [14]. Fig. 2 illustrates the implemented attack. Specifically, an invasion model is designed to generate synthetic data $\mathbf{q}$ from white noise $\mathbf{n}$ derived from a Gaussian distribution. The invasion model consists of several 2D transposed convolution layers which apply fractionally-strided convolution operation on the input to derive a pseudo deconvolution of the input. It takes in the white noise signal $\mathbf{n}$ and outputs a synthetic signal $\mathbf{q}$ of the same dimension as the true EEG signal. The generated data $\mathbf{q}$ is then submitted to the target user verification model to obtain the corresponding confidence score s. A loss function comparing s to 1 (the user class label) is utilized to optimize the parameters of the invasion model. After the training process, the attacker can obtain an invasion model that generates synthetic data to pass the user verification model.

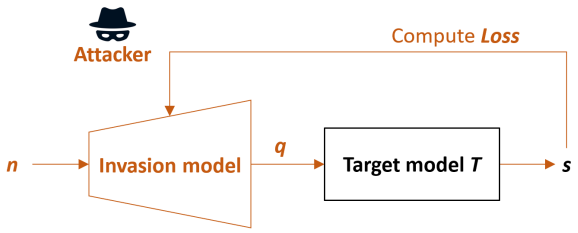


Fig. 2. Model inversion attack on deep learning based EEG biometric system.

C. Evaluation of EEG Verification Model Under Attack

We investigate the security of the state-of-the-art deep learning-based user verification model for EEG biometrics, SOTA-DL, under the implemented attack on two databases, SEED and BED. Details of the databases and signal preprocessing procedures are presented in Section V. The implementation of the SOTA-DL user verification model follows a convolutional neural network architecture which has been demonstrated effective in extracting highly-distinctive features from EEG signals for biometric recognition. Table VI in the *Appendix* provides the network configuration details of the SOTA-DL user verification model. Note that the network configuration needs to fit the number of channels and sampling rate of the data, therefore, adjustment applies for different databases. The training adopts the ADAM optimizer with a learning rate of 0.0001, batch size of 16, training epochs of 600, and a dropout rate of 0.2.

For each subject, we train a SOTA-DL verification model as the target model subject to attacks. Then an invasion model is trained to attack the target model. The neural network configuration for the invasion model is summarized in Table VII in the *Appendix*. We train the invasion model with 1000 epochs, 100 batches per epoch and 100 samples per batch. After the model converged, we generate 1000 test samples using the invasion model to test whether these samples can pass the target user verification model. The success rate is calculated as the number of samples accepted divided by the total number of test samples (1000). Table I summarizes the attack success rate for each subject in the two databases. The results show that the EEG user authentication system based on SOTA-DL is vulnerable to model inversion attack, that is, an attacker can launch such attacks to obtain synthetic data to gain unauthorized access to the system, thereby compromising the entire system.

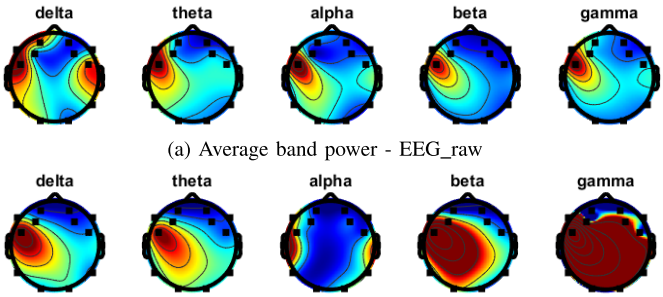
However, having a synthetic data that can pass the authentication model does not necessarily mean that the synthetic data reflect the characteristics of the real raw data from the user. We will refer to the synthetic data generated by invasion model in the model inversion attack as the attack signal, i.e., EEG_attack. To investigate whether EEG_attack is close

TABLE I
SUCCESS RATE (%) OF MODEL INVERSION ATTACKS ON SOTA-DL

SEED (Movie)		*BED (Resting-mixed)*	
Subject	Success Rate	Subject	Success Rate
1	100	1	100
2	100	2	100
3	100	3	100
4	100	4	100
5	100	5	100
6	100	6	100
7	100	7	100
8	100	8	100
9	100	9	100
10	100	10	100
11	100	11	100
12	100	12	100
13	100	13	100
14	100	14	100
15	100	15	100
–	–	16	100
–	–	17	100
–	–	18	100
–	–	19	100
–	–	20	100
–	–	21	100

to the real signal EEG_raw, we perform power spectrum analysis and functional connectivity analysis on these signals and compare their characteristics at the feature level. EEG power and connectivity are two of the most important methods for evaluating EEG signals, because the prominent features of EEG are usually in the frequency domain and manifest as interdependence of signals from different channels [5], [26]. Specifically, Fig. 3 visualizes the average band power of the five canonical EEG frequency bands over the scalp, and Fig. 4 shows the beta band functional connectivity networks (FCNs) computed by phase locking value (PLV). The example is from Subject 2 under EO protocol on BED dataset. We can observe different patterns for the real signal and synthetic attack signal, indicating that the synthetic signal generated by the invasion model cannot reflect the real characteristics of the user's EEG signal. That is to say, although the attack data obtained through model inverse attack can pass through the system, it does not necessarily follow the same data distribution as the private user data. One possible explanation is that the optimization process for the invasion model pushes the generated data to have a confidence score close to 1, which may not always be the case for the target verification model.

In summary, the experimental results demonstrate that model inversion attacks can be utilized to obtain synthetic data to gain unauthorized access to deep learning model-based EEG biometric systems. However, this synthetic data may not reflect the true characteristics of the user's EEG at the feature level. Similar results have been observed in the hill-climbing attack for EEG biometrics with cancellable templates [10], [11], where the attacker launched hill-climbing attacks to obtain synthetic data which can fool the biometric system but is not similar to the genuine user data. To defend against attackers using synthetic data obtained from model inversion attacks to gain unauthorized access, additional security mechanisms need to be designed.



(a) Average band power - EEG_raw

(b) Average band power - EEG_attack

Fig. 3. Visualization of the average band power over the scalp for raw EEG signal (top) and synthetic signal generated by the invasion model in the model inversion attack (bottom) for SOTA-DL. The five canonical EEG frequency bands are considered.

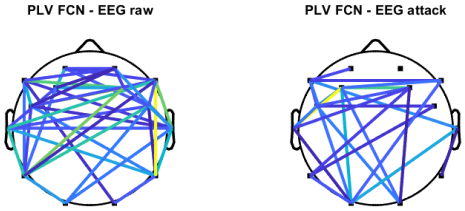


Fig. 4. Visualization of the beta band functional connectivity network using the phase locking value for raw EEG signal (left) and synthetic signal generated by the invasion model in the model inversion attack (right) for SOTA-DL.

IV. PROPOSED METHOD

In this session, we present the design of the cancellable deep learning framework that enhances the security of EEG biometrics, and elaborate the implementation of each component.

A. Cancellable Deep Learning Framework

An overview of our framework is illustrated in Fig. 5. It consists of four major components, including a non-invertible transformation module, a generator, a discriminator and a verification classifier.

During the enrollment stage, raw EEG signals are collected from the user under a pre-defined acquisition protocol (e.g., resting state protocol), and then pre-processed to remove noise and artifacts. Let $\mathbf{x}$ denote the pre-processed EEG data. We convert $\mathbf{x}$ into $\mathbf{y}$ through a non-invertible transform $T(\cdot, k)$ such that $\mathbf{y} = T(\mathbf{x}, k)$, where k is a random key of the transformation to revoke transforms. Changing the value of k can produce different transforms. The non-invertibility property of T guarantees that it is infeasible to invert a given $\mathbf{y}$ and T to get $\mathbf{x}$. Then we train a generative adversarial network (GAN) to approximate the applied transform T. Specifically, the generator takes in $\mathbf{x}$ and outputs $\mathbf{z}$, and the discriminator tries to discriminate the generated data $\mathbf{z}$ from the transformed data $\mathbf{y}$ until they are no longer distinguishable. At this point, we can consider the generator to have successfully replaced the transform. We refer to the training of the generator and

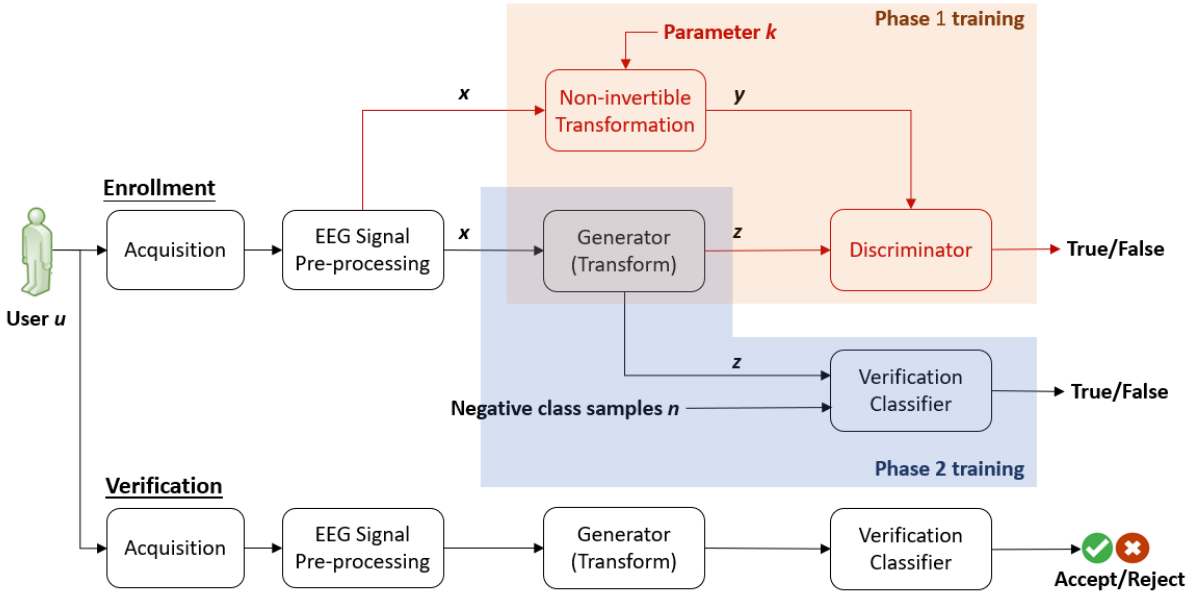


Fig. 5. Proposed cancellable deep learning framework for EEG biometrics. At the end of the enrollment stage, the 'Non-invertible Transformation' module and 'Discriminator' module (red box) are discarded. During the enrollment stage, Phase 1 training jointly updates the Generator and Discriminator, and Phase 2 training updates the Verification Classifier while freezing the Generator obtained in Phase 1.

discriminator as Phase 1 training, where the goal is to obtain a generator that can replace the transformation. In Phase 2 training, a verification classifier is established using data produced from the generator and negative data samples. During this process, the parameters of the generator model are frozen and the optimization only updates the verification classifier. The negative data samples are EEG signals from other users or subjects, and can be collected from open EEG databases. After training, a user model that consists of the generator and verification classifier will be stored in the system, while information about the transformation (including the random key) and discriminator are discarded.

In the verification phase, EEG signals are collected and preprocessed through the same procedures. Then preprocessed data (a probe sample) is fed into the user model (generator and verification classifier) to make a decision whether to accept or reject the request. The framework offers cancellability to the deep learning models stored in the system. If the user model is compromised due to a security breach, we can revoke the compromised model and replace it with a new one. To do this, we randomly produce a new key to renew the transform and repeat the training process to obtain a new user model. The following content of this section will elaborate on each component of the proposed framework.

B. Non-Invertible Transformation

Non-invertible transformation is an effective way to achieve cancellable templates to protect the raw biometrics. It is usually a many-to-one function designed to modify raw biometric data into a new form within the feature or signal space. Implementations of irreversible transformations include algorithms based on random projections [10], [31] and polynomial functions [11], [32]. In this study, we implemented the random projection algorithm used in [10].

Let $\mathbf{x}$ be a EEG data sample, we have $\mathbf{x} \in \mathbb{R}^{N_c \times N_s}$, where N_c and N_s denote the number of EEG channels and the number of EEG time points, respectively. We initialize a random seed k and generate a linear projection matrix $\mathbf{M}$ through a random number generator, where $\mathbf{M}$ is of dimension $N_s \times N_t$ (N_t is significantly smaller than N_s) and contains random scalars drawn from the uniform distribution in the interval $(0, 1)$. Hence, we can transform the input signal into an encoding via the linear projection matrix, i.e., $\mathbf{y} = \mathbf{x} \cdot \mathbf{M}$ and $\mathbf{y} \in \mathbb{R}^{N_c \times N_t}$. Since the projection matrix $\mathbf{M}$ forms an under-determined system of equations, the transformation is non-invertible. To generate a new transform, we only need to change the random seed k and update the projection matrix M.

The above projection process is a popular way to achieve cancellable template designs, however, it only provides one-time-pad security [31] and cannot resist the Attacks via Record Multiplicity (ARM) [33]. The ARM attack is a well-known attack that can retrieve raw biometrics data through compromising and correlating multiple templates. In our framework, sensitive information such as transformed data and non-invertible transformation parameter are not stored in the system, therefore, ARM attack will be infeasible. It is also worth noting that the proposed framework is not confined to specific transformation. Completely different transformation can be applied when revoking a compromised model.

C. Generative Model for Transformation Learning

We use a GAN model to approximate the transform derived from the non-invertible transformation module. With adversarial play between two components, a generator and a discriminator, the GAN learns a function approximation. Specifically, the generator takes $\mathbf{x}$ and learns a distribution

of the target data $\mathbf{y}$ via exploiting the prior distribution of its input $p(\mathbf{x})$ and the generator function $G(\mathbf{x}; \Theta_G)$, where Θ_G is the parameter set of the G function. The discriminator component learns a differentiable function $D(\mathbf{y}; \Theta_D)$, which attempts to distinguish whether an input $\mathbf{y}$ is from the true target distribution $p(\mathbf{y})$ or the generator function $G(\mathbf{x}; \Theta_G)$. The discriminator is trained through minimizing the mean squared error between the predicted label and the true label of each sample, while the generator is trained by minimizing the function $log(1 - D(G(\mathbf{z}; \Theta_G))$. Hence, the adversarial optimization problem is formed as:

$$min_G max_D F(D, G) = \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}[log D(\mathbf{y}; \Theta_D))] \quad (1)$$

$$+ \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[log(1 - D(G(\mathbf{x}; \Theta_G)))] \quad (2)$$

After training the GAN, the discriminator and transformation components are discarded, and only the generator is retained for training the verification classifier.

D. User Verification Model

The verification model adopts the state-of-the-art deep learning model for EEG biometrics. It consists of several convolutional blocks, followed by a fully connected layer to aggregate features and output predictions. The verification classifier is trained using the binary cross entropy loss as follows:

$$Loss = \sum_i -(y_i * log(p_i) + (1 - y_i) * log(1 - p_i)), \quad (3)$$

where y_i represents the actual class label and $log(p_i)$ is the probability of that class.

V. Experimental Evaluation

A. Database and Pre-Processing

EEG data used in our experiments are collected from two databases, including the Biometric EEG dataset (BED) [34] and the SJTU Emotion EEG Dataset (SEED) [35]. The two databases contains EEG recordings of 19 and 15 subjects, respectively. Different EEG elicitation protocols were adopted for signal acquisition, and we select the resting states with eye-open (EO) and eye-closed (EC) from BED since resting states have been demonstrated to be effective and convenient protocols for EEG biometrics [36]. The SEED is an emotion database that focuses on the EEG of subjects watching movie clips, where the movie clips are designed to trigger positive, negative and neutral emotions in the subjects. Considering that emotion may have an impact on biometric recognition, we select EEG under the neutral condition for our experiment. Moreover, the two databases utilized different EEG acquisition equipment, including a medical-grade system (NeuroScan) and a consumer-grade device (Emotive EPOC+), hence the signal quality, number of channels and signal sampling rate vary. In addition, both databases provide three sessions of recordings, allowing the evaluation of cross-session performance. Details about the databases are summarized in Table II.

TABLE II
Databases

Datasets	#Subj.	#Ch.	#Sess.	SamplingRate	Protocols	Devices
BED	21	14	3	256 Hz	EO EC	EPOC+**
SEED	15	62	3	200 Hz	Movie (neutral)	NeuroScan*

*medical-grade **consumer-grade

Raw EEG signals collected from the sensors are usually contaminated with noise and artifacts, hence we perform an automatic EEG signal pre-processing pipeline, the HAPPE [37], for denoising and artifact removal. Specific pre-processing steps include filtering (alpha and beta bands [38]), bad channel interpolation, ICA and artifact component rejection (with the MARA algorithm [39]), and re-referencing (common average referencing). The signal is also downsampled to half its original sampling rate for efficiency. Finally, the preprocessed signal is segmented into two-second samples by a non-overlapping moving window. Therefore, each sample contains two seconds of EEG signals which are 14×256 and 62×200 time points for data in BED and SEED, respectively.

B. Comparison Methods and Evaluation Metrics

The proposed method is compared with the state-of-the-art algorithms for EEG biometric verification, including:

- State-of-the-art template based method for EEG biometrics (SOTA-Temp). The template consists of EEG autoregressive features, power spectral features, fuzzy entropy features, and graph features calculated from the EEG functional connectivity networks [5], [8].
- Cancellable EEG template design (Cancellable-Temp) [11]. This method transforms the EEG graph features into the encrypted domain through a non-invertible transformation based on polynomial equations.
- State-of-the-art deep learning model based on convolutional neural networks for EEG biometrics (SOTA-DL) [12].

For verification, we report the classification accuracy and equal error rate (EER) which is the error rate when the false match rate (FMR) is equal to the false non-match rate (FNMR). A false match happens when a non-user sample is misclassified as user by the verification model, and a false non-match is when a user sample is incorrectly recognized as impostor. In addition, the detection error trade-off (DET) curves are used to show the tradeoff between FMR and FNMR.

C. Network and Training Configurations

As suggested by existing findings, the implementation of neural network architectures should take into account the characteristics of specific signals. We slightly adjust the neural network configurations for signals from different databases considering the number of channels and sampling rate. The configuration details are summarized in Table VIII in the

TABLE III
WITHIN-SESSION VERIFICATION ACCURACY (%) AND EER (%) OF THE PROPOSED METHOD AND COMPARISON METHODS ON TWO DATABASES

Method	Performance	BED			SEED
		EO	EC	Resting-mixed	Movie (neutral)
SOTA-Temp	Acc	91.1	87.98	84.24	95.92
	EER	8.94	12	15.78	4.03
Cancellable-Temp	Acc	87.31	87.65	81.59	95.3
	EER	12.7	12.36	18.41	4.71
SOTA-DL	Acc	93.72	90.13	95.26	90.11
	EER	6.28	9.87	4.74	9.89
Cancellable-DL (proposed)	Acc	95.87	96.86	97.04	97.88
	EER	4.13	3.14	2.96	2.12

TABLE IV
CROSS-SESSION VERIFICATION ACCURACY (%) AND EER (%) OF THE PROPOSED METHOD AND COMPARISON METHODS ON TWO DATABASES

Method	Performance	BED			SEED
		EO	EC	Resting-mixed	Movie (neutral)
SOTA-Temp	Acc	62.63	64.49	59.21	61.03
	EER	37.28	35.64	40.66	38.97
Cancellable-Temp	Acc	61.10	63.10	57.24	62.28
	EER	38.94	37.91	42.83	37.71
SOTA-DL	Acc	65.52	73.31	70.47	64.87
	EER	34.48	26.69	29.53	35.13
Cancellable-DL (proposed)	Acc	71.12	74.79	71.41	71.69
	EER	28.88	25.21	28.59	28.31

Appendix. To have a fair comparison, the user verification model (classifier) in the proposed framework shares the same architecture and settings as the SOTA-DL model. For training, the ADAM algorithm is adopted for stochastic gradient optimization with a learning rate of 0.0001, batch size of 16, training epochs of 600, and a dropout rate of 0.2.

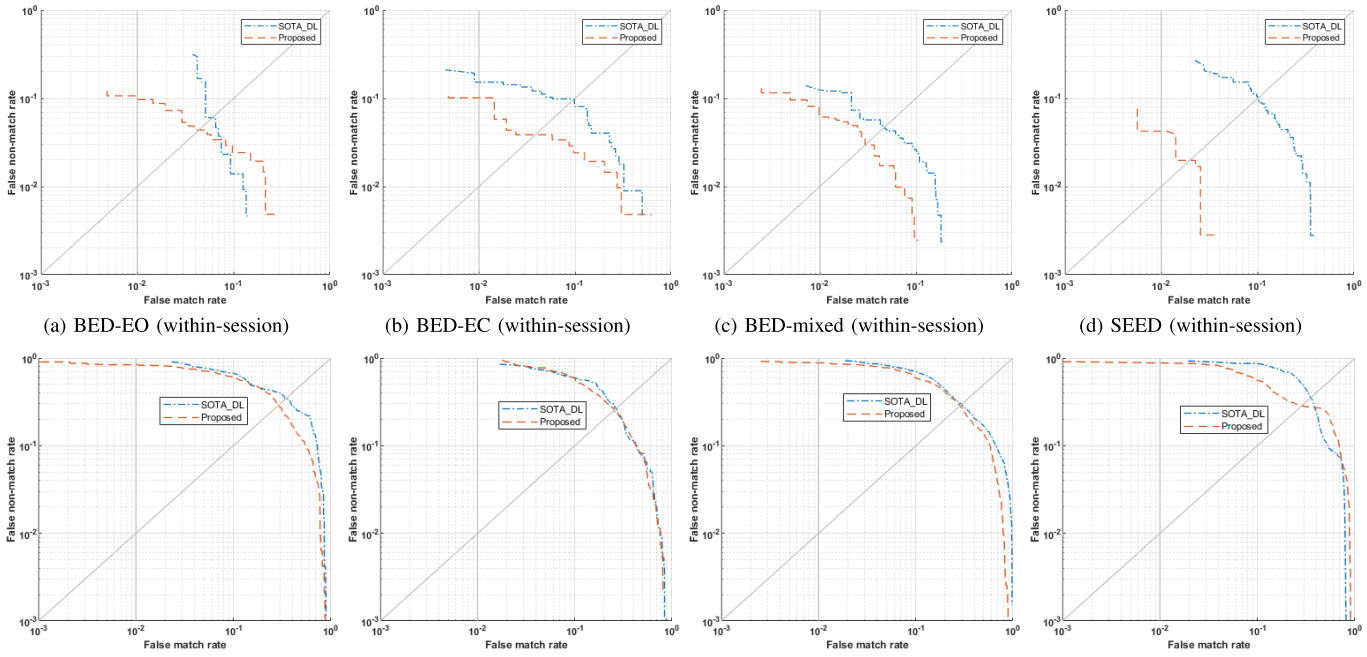
We perform both within-session evaluation and cross-session evaluation. In the within-session experiment, both the training and testing are performed on session 1 data. To ensure that no data from any test impostor can be seen by the model during the training stage, we separate the importer set from the user set. Specifically, for each user, we separate the subsequent five consecutive subjects as the impostor test set. Then, 80% of the user data and the same amount of data randomly drawn from the other users are used for training the networks. During the testing stage, the remaining 20% user data and all data from the impostor set are used for testing the verification performance. The cross-session evaluation follows a similar procedure, except that the training uses session 1 and 2 data, and the testing is performed on session 3 data. The source code will be available online.[1]

D. Verification Performance

Table III presents the within-session verification performance of the proposed method and comparison methods in terms of classification accuracy and EER. We can see that deep learning model-based methods achieved significantly better performance compared to the template-based approaches in most of the cases. This improvement demonstrates the capability of deep neural network models in learning effective identity-bearing representations from the raw signals for accurate user verification. It also explains the increasing popularity of deep learning methods in EEG classification for brain-computer interfaces. Therefore, designing security mechanisms for deep learning-based EEG verification systems is an important topic worthy of studying. Comparing results of SOTA-Temp and Cancellable-Temp, it can be seen that the use of cancellable template design (non-invertible transformation) degraded the verification performance to varying degrees. This often happens because non-invertible transformation resets the order or position of the feature set, enlarging intra-class variations hence weakening the discriminatory power of the transformed features. In comparison, the proposed Cancellable-DL achieved better performance than the corresponding SOTA-DL which is non-cancellable. In other words, the proposed framework grants cancelability without compromising the verification performance. This is because the generator in our framework is user-specific and only learns to transfer data distribution of the genuine user, hence impostor data transformed by the generator can be better differentiated from the user data distribution.

The cross-session performance is provided in Table IV. Due to the intra-person variation of EEG signals, all methods show performance degradation in cross-session experiments. Similarly, the proposed cancellable deep learning method achieved better verification performance than template-based approaches, including non-cancellable and cancellable ones. It also outperformed the state-of-the-art deep learning model based method, which is non-cancellable. Fig. 6 shows

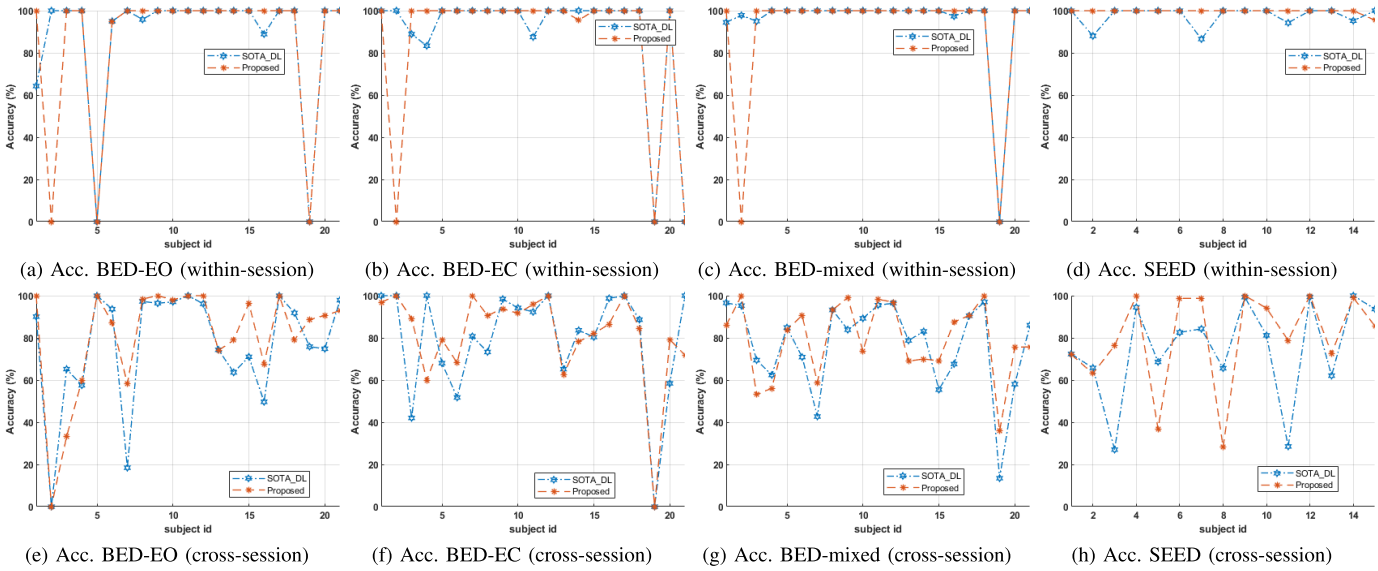
[1] https://github.com/HubYZ/CancelableEEGDeepLearning



(a) BED-EO (within-session) (b) BED-EC (within-session) (c) BED-mixed (within-session) (d) SEED (within-session)

(e) BED-EO (cross-session) (f) BED-EC (cross-session) (g) BED-mixed (cross-session) (h) SEED (cross-session)

Fig. 6. Comparison of DET curves obtained by SOTA-DL and Cancellable-DL (proposed) in within-session and cross-session evaluation.



(a) Acc. BED-EO (within-session) (b) Acc. BED-EC (within-session) (c) Acc. BED-mixed (within-session) (d) Acc. SEED (within-session)

(e) Acc. BED-EO (cross-session) (f) Acc. BED-EC (cross-session) (g) Acc. BED-mixed (cross-session) (h) Acc. SEED (cross-session)

Fig. 7. Verification accuracy of SOTA-DL and Cancellable-DL (proposed) for each subject in within-session and cross-session evaluation.

the corresponding DET curves obtained by SOTA-DL and Cancellable-DL in the within-session and cross-session evaluation.

In addition, we also report the verification accuracy of SOTA-DL and Cancellable-DL for each subject in Fig. 7. We can notice that some subjects are difficult to verify due to the considerable large variations of the signals. This observation is consistent with previous studies and poses an interesting question for the longitudinal analysis of EEG biometrics [40]. In summary, the proposed method provides cancellability to verification models while providing better performance than the state-of-the-art deep learning models for EEG user verification.

E. Security Under Model Inversion Attack

In Section III, we perform the model inversion attack on SOTA-DL, and demonstrate that the state-of-the-art deep learning-based EEG user verification model is vulnerable to the attack, with 100% attack success rate for all subjects in BED and SEED datasets. In this section, we investigate whether the proposed cancellable deep learning framework can resist the second attack by revoking the compromised model. Fig. 8 depicts the second attack [10], which refers to the use of data generated from the invasion model derived from model inversion attacks to try to break into the system after the system has revoked the compromised models. Note that the second attack is a concept for cancellable biometrics. For

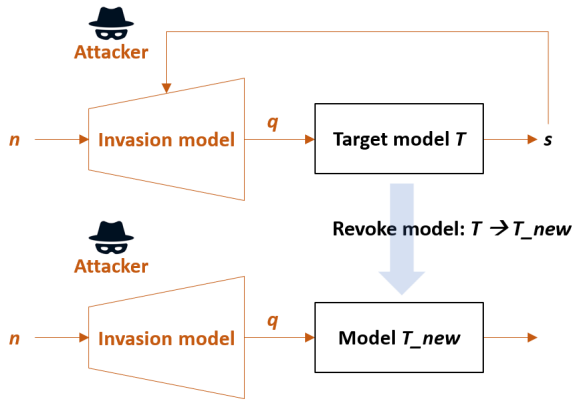


Fig. 8. Model inversion attack and second attack.

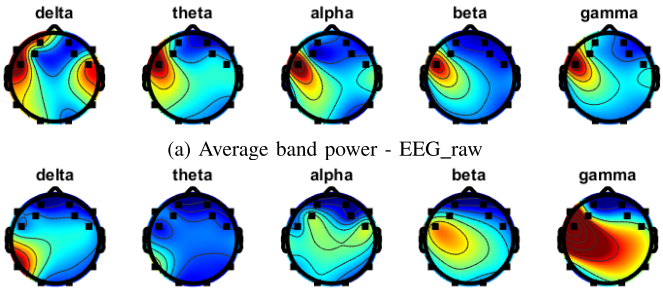
TABLE V

SUCCESS RATE (%) OF THE MODEL INVERSION ATTACK AND SECOND ATTACK FOR THE PROPOSED METHOD

SEED (Movie)			*BED (Resting-mixed)*		
Subject	Attack Success Rate		Subject	Attack Success Rate	
	First	Second		First	Second
1	100	0	1	100	0
2	100	0	2	100	100
3	100	0	3	100	0
4	100	6.4	4	100	99.8
5	100	0	5	100	0
6	100	0	6	100	0
7	100	0	7	100	0
8	100	0	8	100	0
9	100	100	9	100	100
10	100	0	10	100	0
11	100	0	11	100	0
12	100	0	12	100	0
13	100	0	13	100	0
14	100	73.7	14	100	0
15	100	0	15	100	0
–	–	–	16	100	0
–	–	–	17	100	0
–	–	–	18	100	0.2
–	–	–	19	100	0
–	–	–	20	100	0
–	–	–	21	100	0

non-cancellable methods such as SOTA-DL, the entire system is already compromised after the first attack.

Success rates of the model inversion attack and second attack for each subject in SEED and BED are summarized in Table V. We can see that the success rate of the model invasion attack (first attack) is 100% for all subjects, meaning that the adversary is able to find synthetic samples that can pass the user verification model. This observation is consistent with the one for SOTA-DL in Section III, showing that model inversion attacks using generative networks poses a threat to model-based EEG biometric systems. Specifically, an adversary can optimize an invasion model and exploit the confidence scores produced by the target model to generate a synthetic sample to gain false acceptance, thereby compromising the target model.



(a) Average band power - EEG_raw

(b) Average band power - EEG_attack

Fig. 9. Visualization of the average band power over the scalp for raw EEG signal (top) and synthetic signal generated by the invasion model in the model inversion attack (bottom) for the proposed method Cancellable-DL. The five canonical EEG frequency bands are considered.

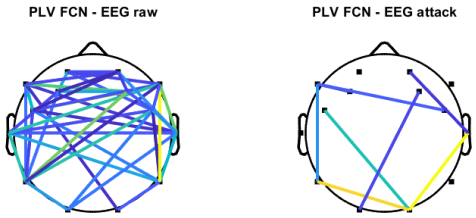


Fig. 10. Visualization of the beta band functional connectivity network using the phase locking value for raw EEG signal (left) and synthetic signal generated by the invasion model in the model inversion attack (right) for the proposed method Cancellable-DL.

Fig. 9 and Fig. 10 visualize the average band power over the scalp and beta band PLV FCNs for raw EEG signal and attack signal generated by the invasion model, respectively. The example is from Subject 2 under EO protocol on BED dataset. We can observe that the real EEG signals from the user and the attack signals generated by the invasion model exhibit different patterns in terms of power features and functional connectivity features, indicating a clear dissimilarity between the two signals. The finding is consistent with that from SOTA-DL that the synthetic signal generated by the invasion model cannot reflect the real characteristics of the user's EEG signal. That is to say, although the attack data obtained through model inverse attack can pass through the system, it does not necessarily follow the same data distribution as the private user data.

In previous non-cancellable systems, once a synthetic data sample is found for the target model, the entire user verification system is compromised. Our proposed framework offers cancellability competencies to the user verification model: the compromised model can be revoked and a new one can be issued by changing the transformation parameters. Table V shows that the success rate of the second attack drops significantly to zero for most users, indicating the effectiveness of the proposed method in enhancing system security and re-usability. The second attack results also reflect the discrepancies between the data distribution learned by the invasion model and the real user data distribution.

Overall, the experimental results demonstrate that the proposed cancellable deep learning framework significantly reduces the possibility of attackers exploiting synthetic data obtained by model inversion attacks to gain unauthorized

TABLE VI
USER VERIFICATION MODEL (SOTA-DL) CONFIGURATION

SEED Database:

Classifier Layers	Kernel Size	Out Ch.	MaxPool	Stride	Padding	BN/AF/DP
$Conv_1$	3×3	128	[2,2]	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_2$	3×3	256	[2,2]	[1,1]	[0, 1]	Yes/ELU/Yes
$Conv_3$	3×3	512	[2,2]	[1,1]	[0, 1]	Yes/ELU/Yes
$Conv_4$	3×3	1024	[2,2]	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_5$	2×9	1024	nul	[1,1]	[0, 0]	Yes/ELU/Yes
FL	nul	nul	nul	nul	nul	No/Sigmoid/No

BED Database:

Classifier Layers	Kernel Size	Out Ch.	MaxPool	Stride	Padding	BN/AF/DP
$Conv_1$	3×3	128	[2,2]	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_2$	3×3	256	[2,2]	[1,1]	[0, 1]	Yes/ELU/Yes
$Conv_3$	2×4	512	[1,4]	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_4$	1×13	1024	nul	[1,1]	[0, 0]	Yes/ELU/Yes
FL	nul	nul	nul	nul	nul	No/Sigmoid/No

BN: Batch normalization; **AF**: Activation function; **DP**: Dropout; **FL**: Fully connected layer

TABLE VII
INVASION MODEL CONFIGURATION

SEED Database:

Invasion Layers	Kernel Size	Out Ch.	MaxPool	Stride	Padding	BN/AF/DP
$DeConv_1$	2×5	1024	nul	[1,1]	[0, 0]	Yes/ELU/Yes
$DeConv_2$	2×4	512	nul	[2,2]	[0, 1]	Yes/ELU/Yes
$DeConv_3$	2×4	256	nul	[2,2]	[0, 1]	Yes/ELU/Yes
$DeConv_4$	3×6	128	nul	[2,1]	[1, 0]	Yes/ELU/Yes
$DeConv_5$	2×4	64	nul	[2,2]	[0, 1]	Yes/ELU/Yes
$DeConv_6$	2×4	32	nul	[2,2]	[0, 1]	Yes/ELU/Yes
$DeConv_7$	3×4	1	nul	[1,2]	[0, 1]	No/No/No

BED Database:

Invasion Layers	Kernel Size	Out Ch.	MaxPool	Stride	Padding	BN/AF/DP
$DeConv_1$	1×4	1024	nul	[1,1]	[0, 0]	Yes/ELU/Yes
$DeConv_2$	2×4	512	nul	[2,2]	[0, 1]	Yes/ELU/Yes
$DeConv_3$	2×4	256	nul	[2,2]	[0, 1]	Yes/ELU/Yes
$DeConv_4$	2×4	128	nul	[2,2]	[0, 1]	Yes/ELU/Yes
$DeConv_5$	2×4	64	nul	[2,2]	[3, 1]	Yes/ELU/Yes
$DeConv_6$	3×4	32	nul	[1,2]	[0, 1]	Yes/ELU/Yes
$DeConv_7$	3×4	1	nul	[1,2]	[0, 1]	No/No/No

BN: Batch normalization; **AF**: Activation function; **DP**: Dropout

TABLE VIII
MODEL CONFIGURATION FOR GENERATOR, DISCRIMINATOR AND CLASSIFIER OF THE PROPOSED CANCELLABLE DEEP LEARNING FRAMEWORK

SEED Database:

Generator Layers	Kernel Size	Out Ch.	MaxPool	Stride	Padding	BN/AF/DP
$Conv_1$	1×9	64	nul	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_2$	1×9	64	nul	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_3$	1×9	64	nul	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_4$	1×9	62	nul	[1,1]	[0, 0]	Yes/No/No
$Conv_5$	1×1	1	nul	[1,1]	[0, 0]	Yes/No/No
Discriminator Layers	**Kernel Size**	**Out Ch.**	**MaxPool**	**Stride**	**Padding**	**BN/AF/DP**
$Conv_1$	1×4	64	[1,2]	[1,2]	[0, 1]	Yes/ELU/Yes
$Conv_2$	1×4	64	[1,2]	[1,2]	[0, 1]	Yes/ELU/Yes
$Conv_3$	1×4	64	nul	[1,2]	[0, 1]	Yes/ELU/Yes
$Conv_4$	1×5	64	nul	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_5$	62×1	1	nul	[1,1]	[0, 0]	No/Sigmoid/No
Classifier Layers	**Kernel Size**	**Out Ch.**	**MaxPool**	**Stride**	**Padding**	**BN/AF/DP**
$Conv_1$	3×3	128	[2,2]	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_2$	3×3	256	[2,2]	[1,1]	[0, 1]	Yes/ELU/Yes
$Conv_3$	3×3	512	[2,2]	[1,1]	[0, 1]	Yes/ELU/Yes
$Conv_4$	3×3	1024	[2,2]	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_5$	2×9	1024	nul	[1,1]	[0, 0]	Yes/ELU/Yes
FL	nul	nul	nul	nul	nul	No/Sigmoid/No

BED Database:

Generator Layers	Kernel Size	Out Ch.	MaxPool	Stride	Padding	BN/AF/DP
$Conv_1$	1×9	64	nul	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_2$	1×9	64	nul	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_3$	1×9	64	nul	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_4$	1×9	14	nul	[1,1]	[0, 0]	Yes/No/No
$Conv_5$	1×1	1	nul	[1,1]	[0, 0]	Yes/No/No
Discriminator Layers	**Kernel Size**	**Out Ch.**	**MaxPool**	**Stride**	**Padding**	**BN/AF/DP**
$Conv_1$	1×4	64	[1,2]	[1,2]	[0, 1]	Yes/ELU/Yes
$Conv_2$	1×4	64	[1,2]	[1,2]	[0, 1]	Yes/ELU/Yes
$Conv_3$	1×4	64	nul	[1,2]	[0, 1]	Yes/ELU/Yes
$Conv_4$	1×7	64	nul	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_5$	14×1	1	nul	[1,1]	[0, 0]	No/Sigmoid/No
Classifier Layers	**Kernel Size**	**Out Ch.**	**MaxPool**	**Stride**	**Padding**	**BN/AF/DP**
$Conv_1$	3×3	128	[2,2]	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_2$	3×3	256	[2,2]	[1,1]	[0, 1]	Yes/ELU/Yes
$Conv_3$	2×4	512	[1,4]	[1,1]	[0, 0]	Yes/ELU/Yes
$Conv_4$	1×13	1024	nul	[1,1]	[0, 0]	Yes/ELU/Yes
FL	nul	nul	nul	nul	nul	No/Sigmoid/No

access to EEG biometric systems. The cancellability mechanism enhances the system security and re-usability. We also notice that the second attack rate for certain subjects (e.g., subject 9, 14 in SEED and subject 2, 4, 8 in BED) is high. This may be an issue related to intra-person variation [2], [4] and the trade-off between prediction and security [14]. A recent finding reveals a paradoxical relationship between a model's predictive power and its susceptibility to general model inversion attacks, that is, models with stronger predictive power are more sensitive to the inversion attacks [14]. We will dig into this issue in our future study.

VI. CONCLUSION

Deep learning-based EEG biometric systems train and store a predictive model for each user in the system for verification. We demonstrate that such deep learning model-based biometric systems are vulnerable to malicious attacks to gain unauthorized access. Our results show that it is feasible for an imposter to exploit a generative model to obtain synthetic data to pass the biometric system. In such cases, the verification model is compromised, and the corresponding biometric system is no longer secure. In this paper, we present a cancellable deep learning framework to protect deep learning model-based EEG biometric systems. The framework consists of a generator, a discriminator, a non-invertible transformation, and a verification classifier. During the enrollment stage, EEG data of the user are transformed through the non-invertible transformation, and we jointly train the generator and discriminator in a GAN structure to approximate the non-invertible transformation. Then a verification classifier is established using data produced by the generator. After training, a predictive model that consists of the generator and verification classifier will be stored in the system, while information about the transformation (including the random seed) and discriminator are discarded. When a model is compromised by model inversion attacks or hill-climbing attacks, the framework is able to issue a new transformation and produce a new predictive model to replace the compromised one. Our experimental results show that our proposed method provides comparative verification accuracy while offering model cancellability competency. In our future study, we will look into the intra-subject variation

problem of EEG and aim to improve cross-session verification performance, which is an open research question. In addition, a deep learning-based EEG user verification model may leak private information (signal characteristics or features) through its hidden layers in a white-box setting, which means the attacker can infer important physical traits of a user's EEG through the model. The proposed cancellable deep learning framework can provide additional security capabilities to protect user privacy since models trained on transformed data do not reflect physically interpretable features. The privacy issue of deep learning-based EEG biometrics is still an open research question, and we will investigate it in our future work.

APPENDIX

Details about the neural network configuration used in the experiments are summarized in Table VII and Table VIII.

REFERENCES

[1] M. Wang, H. El-Fiqi, J. Hu, and H. A. Abbass, "Convolutional neural networks using dynamic functional connectivity for EEG-based person identification in diverse human states," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 12, pp. 3259–3272, Dec. 2019.

[2] Q. Gui, M. V. Ruiz-Blondet, S. Laszlo, and Z. Jin, "A survey on brain biometrics," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–38, Feb. 2019.

[3] M. V. Ruiz-Blondet, Z. Jin, and S. Laszlo, "CEREBRE: A novel method for very high accuracy event-related potential biometric identification," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1618–1629, Jul. 2016.

[4] M. Wang, X. Yin, Y. Zhu, and J. Hu, "Representation learning and pattern recognition in cognitive biometrics: A survey," *Sensors*, vol. 22, no. 14, p. 5111, Jul. 2022.

[5] M. Wang, J. Hu, and H. A. Abbass, "BrainPrint: EEG biometric identification based on analyzing brain connectivity graphs," *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107381.

[6] S. Yang, F. Deravi, and S. Hoque, "Task sensitivity in EEG biometric recognition," *Pattern Anal. Appl.*, vol. 21, no. 1, pp. 105–117, Feb. 2018.

[7] Y. Höller and A. Uhl, "Do EEG-biometric templates threaten user privacy?" in *Proc. 6th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2018, pp. 31–42.

[8] C. He, X. Lv, and Z. Jane Wang, "Hashing the MAR coefficients from EEG data for person authentication," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 1445–1448.

[9] R. Damaševič ius, R. Maskeliūnas, E. Kazanavič ius, and M. Woźniak, "Combining cryptography with EEG biometrics," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–11, Jan. 2018.

[10] M. Wang, S. Wang, and J. Hu, "Cancellable template design for privacy-preserving EEG biometric authentication systems," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3350–3364, 2022.

[11] M. Wang, S. Wang, and J. Hu, "PolyCosGraph: A privacy-preserving cancelable EEG biometric system," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 5, pp. 4258–4272, Sep./Oct. 2022.

[12] A. J. Bidgoly, H. J. Bidgoly, and Z. Arezoumand, "Towards a universal and privacy preserving EEG-based authentication system," *Sci. Rep.*, vol. 12, no. 1, pp. 1–12, Feb. 2022.

[13] T. Zhu, D. Ye, S. Zhou, B. Liu, and W. Zhou, "Label-only model inversion attacks: Attack with the least information," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 991–1005, 2023.

[14] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 253–261.

[15] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, "Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 357–372, 2022.

[16] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.

[17] G. Han, J. Choi, H. Lee, and J. Kim, "Reinforcement learning-based black-box model inversion attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20504–20513.

[18] E. Maiorana, G. E. Hine, and P. Campisi, "Hill-climbing attacks on multibiometrics recognition systems," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 900–915, May 2015.

[19] K. Pizzi, F. Boenisch, U. Sahin, and K. Böttinger, "Introducing model inversion attacks on automatic speaker recognition," 2023, *arXiv:2301.03206*.

[20] P. Campisi et al., "Brain waves based user recognition using the 'eyes closed resting conditions' protocol," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, Dec. 2011, pp. 1–6.

[21] S. Yang and F. Deravi, "On the usability of electroencephalographic signals for biometric recognition: A survey," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 6, pp. 958–969, Dec. 2017.

[22] J. Chuang, H. Nguyen, C. Wang, and B. Johnson, "I think, therefore I am: Usability and security of authentication using brainwaves," in *Proc. Int. Conf. Financial Cryptogr. Data Secur.* Berlin, Germany: Springer, 2013, pp. 1–16.

[23] E. Maiorana, D. La Rocca, and P. Campisi, "On the permanence of EEG signals for biometric recognition," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 1, pp. 163–175, Jan. 2016.

[24] Z. Cao and C.-T. Lin, "Inherent fuzzy entropy for the improvement of EEG complexity evaluation," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 1032–1035, Apr. 2018.

[25] M. Fraschini, S. M. Pani, L. Didaci, and G. L. Marcialis, "Robustness of functional connectivity metrics for EEG-based personal identification over task-induced intra-class and inter-class variations," *Pattern Recognit. Lett.*, vol. 125, pp. 49–54, Jul. 2019.

[26] D. L. Rocca et al., "Human brain distinctiveness based on EEG spectral coherence connectivity," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 9, pp. 2406–2412, Sep. 2014.

[27] E. Debie, N. Moustafa, and A. Vasilakos, "Session invariant EEG signatures using elicitation protocol fusion and convolutional neural network," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 4, pp. 2488–2500, Jul. 2022.

[28] E. Maiorana, G. E. Hine, D. L. Rocca, and P. Campisi, "On the vulnerability of an EEG-based biometric system to hill-climbing attacks algorithms' comparison and possible countermeasures," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2013, pp. 1–6.

[29] E. Maiorana, D. L. Rocca, and P. Campisi, "Cognitive biometric cryptosystems a case study on EEG," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Sep. 2015, pp. 125–128.

[30] M. Kahla, S. Chen, H. A. Just, and R. Jia, "Label-only model inversion attacks via boundary repulsion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15025–15033.

[31] S. Wang and J. Hu, "Alignment-free cancelable fingerprint template design: A densely infinite-to-one mapping (DITOM) approach," *Pattern Recognit.*, vol. 45, no. 12, pp. 4129–4137, Dec. 2012.

[32] Q. N. Tran and J. Hu, "A multi-filter fingerprint matching framework for cancelable template design," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2926–2940, 2021.

[33] C. Li and J. Hu, "Attacks via record multiplicity on cancelable biometrics templates," *Concurrency Comput., Pract. Exper.*, vol. 26, no. 8, pp. 1593–1605, Jun. 2014.

[34] P. Arnau-González, S. Katsigiannis, M. Arevalillo-Herráez, and N. Ramzan, "BED: A new data set for EEG-based biometrics," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 12219–12230, Aug. 2021.

[35] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2018.

[36] D. La Rocca, P. Campisi, and G. Scarano, "EEG biometrics for individual recognition in resting state with closed eyes," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2012, pp. 1–12.

[37] L. J. Gabard-Durnam, A. S. M. Leal, C. L. Wilkinson, and A. R. Levin, "The Harvard automated processing pipeline for electroencephalography (HAPPE): Standardized processing software for developmental and high-artifact data," *Frontiers Neurosci.*, vol. 12, p. 97, Feb. 2018.

[38] E. Maiorana, "Learning deep features for task-independent EEG-based biometric verification," *Pattern Recognit. Lett.*, vol. 143, pp. 122–129, Mar. 2021.
[39] I. Winkler, S. Brandl, F. Horn, E. Waldburger, C. Allefeld, and M. Tangermann, "Robust artifactual independent component classification for BCI practitioners," *J. Neural Eng.*, vol. 11, no. 3, Jun. 2014, Art. no. 035013.
[40] E. Maiorana and P. Campisi, "Longitudinal evaluation of EEG-based biometric recognition," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 5, pp. 1123–1138, May 2018.

Min Wang (Member, IEEE) received the Ph.D. degree in computer science from the University of New South Wales, Canberra, Australia. She is currently a Lecturer with the School of Information Technology and Systems, University of Canberra, and an Adjunct Lecturer with the School of Systems and Computing, University of New South Wales. She has published articles in top journals, including IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON CYBERNETICS, and *Pattern Recognition*. Her research interests include biometrics, pattern recognition, privacy and security, and bio-cryptography.

Xuefei Yin received the B.S. degree from Liaoning University, Liaoning, China, the M.E. degree from Tianjin University, Tianjin, China, and the Ph.D. degree from the University of New South Wales Canberra at ADFA, Canberra, Australia. He is currently with the School of Information and Communication Technology, Griffith University, Gold Coast, QLD, Australia. He has published articles in top journals, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, *ACM Computing Surveys*, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, and IEEE INTERNET OF THINGS JOURNAL. His research interests include biometrics, pattern recognition, privacy-preserving, and intrusion detection.

Jiankun Hu (Senior Member, IEEE) is currently a Professor with the School of Engineering and Information Technology, University of New South Wales, Canberra, Australia. He is also an invited Expert of Australia Attorney-General's Office, assisting the draft of Australia National Identity Management Policy. He has received nine Australian Research Council (ARC) Grants and has served at the Panel on Mathematics, Information, and Computing Sciences, Australian Research Council ERA—The Excellence in Research for Australia Evaluation Committee in 2012. His research interests include cyber security covering intrusion detection, sensor key management, and biometrics authentication. His main research interests include cyber security, including biometrics security, where he has publications at top venues, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and *Pattern Recognition*. He is a Senior Area Editor of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.