Task-Preserving EEG Anonymization Using Latent Feature Masking

Nicholas Mehlman, Shrikanth Narayanan

***Abstract*— The growing use of electroencephalogram (EEG) data in deep learning offers the potential for revolutionary new applications while also raising significant privacy concerns. Given the heavily idiosyncratic nature of EEG data, a bad actor might easily exploit public datasets to identify specific subjects and, hence, infer potentially sensitive information about their mental state, health status, etc. In the paper, we propose , a method to selectively remove identity features from EEG data, while preserving the utility of the dataset for its intended purpose. We test SAFE against three distinct threat models using four popular EEG classification datasets. Results demonstrate that SAFE provides strong protection against adversarial actors without seriously compromising task-relevant features. Our approach addresses the important yet largely unexplored need for robust protections and safety against privacy threats to EEG data while supporting data sharing, enabling scientific and clinical advances.**

***Index Terms*— Data Anonymization, Deep Learning, Privacy**

I. Introduction

Recent breakthroughs in machine learning have enabled novel insights into human brain processes in healthy and disease/disorder conditions from electroencephalogram (EEG) data. Deep learning models have demonstrated remarkable performance on a diverse set of tasks, such as emotion recognition [1], [2], seizure detection [3], [4], and motor imagery [5], [6]. While these advances offer the potential for exciting new diagnostic, therapeutic and scientific applications, the use and proliferation of EEG data, including public EEG corpora, also raise serious privacy concerns. In particular, EEG signals contain strong biometric markers [7]–[9] (i.e., features linked to a subject's identity) that a malicious actor could exploit by using these datasets to identify individuals, or misuse in other ways. This threat is compounded by the sensitive information that can be extracted from these data, including mental health status [10], [11], neurological condition [12], [13], and even substance use history [14].

Despite these privacy risks, relatively little attention has been given to developing robust methods of anonymizing EEG. Furthermore, of the prior works that do address this issue (e.g., [15]–[17]), most fail to articulate a well-defined and realistic privacy threat model. In this work, we propose **S**ubject-**A**nonymized **F**unctional **E**EG (SAFE), a novel approach for anonymizing EEG datasets without distorting the information required to perform the intended task (e.g., motor imagery). SAFE uses latent-space masking to selectively suppress features associated with subject identity while preserving those that are task-relevant. To evaluate our method, we define three plausible threat models under which a bad actor might misuse identity information. The first attack, which we term 'unmasking', uses a pre-trained subject classifier to identify (unmask) subjects within the target dataset. The second 'inference' attacker trains a subject classifier on the target dataset with the intent to apply the trained model to infer the identity of individuals in a separate unprotected collection of EEG recordings. The final 'open eval' attacker is assumed to have black-box access to the anonymization system, which they leverage during model training (inference-like) or evaluation (unmasking-like) to reduce the distributional shift between the data they train on and the data that they wish to de-anonymize. Our contributions are as follows:

1) We define a rigorous theoretical framework for anonymization based on information theory
2) We define three practical threat models by which an attacker might exploit subject identity information in open-source EEG datasets
3) We design and test SAFE, a novel method for anonymizing EEG data, and demonstrate empirically that it successfully defends against the proposed threat models when tested on real-world EEG datasets

II. Prior Work

EEG has been shown to encode strong markers of individual identity [7] along with other sensitive attributes such as emotional state [18] and mental health status [19]. Works such as [8] and [9] have demonstrated that it can be used for biometric

This publication was supported by a Subagreement from The Johns Hopkins University with funds provided by sponsor award number 140D0424C0067 from the Office of the Director of National Intelligence (ODNI) through the Department of the Interior, Interior Business Center Acquisition Services Directorate. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Office of the Director of National Intelligence (ODNI), the Department of the Interior, or The Johns Hopkins University.

Nicholas Mehlman is a PhD Student in the Viterbi School of Engineering at the University of Southern California. He is a member of the Signal Analysis and Interpretation Laboratory (SAIL). (e-mail: nmehlman@usc.edu).

Shrikanth Narayanan is a professor in the Signal and Image Processing Institute of the University of Southern California's Ming Hsieh Electrical & Computer Engineering department with joint appointments as Professor in Computer Science, Linguistics, Psychology, Neuroscience, Pediatrics and Otolaryngology-Head and Neck Surgery. He leads the SAIL Laboratory.

identification with a high degree of reliability. Given this, several methods have been proposed for anonymizing EEG data while retaining the information necessary to perform a specific, desirable task such as seizure detection, motor imagery, or sleep stage classification. For example, in [16], the authors proposed introducing small adversarial perturbations that are strongly correlated with subject labels. This misleads identity classifiers by causing them to overfit to the perturbations and hence fail to generalize. In [15], the authors anonymized EEG with an autoencoder-style network along with subject and task regularizers that minimized the reconstruction of identity information while retaining task-relevant features. Several approaches, such as [20] and [17], circumvented the privacy concerns associated with using real EEG data by employing generative adversarial networks (GANs) to create synthetic examples that still encode task-relevant information. GANs were also employed for privacy protection in [21], but on electrocardiogram (ECG) rather than EEG data. Finally, several works (e.g. [22], [23]) have introduced cryptographic approaches for securing EEG data.

While not directly intended for privacy protection, several parallel lines of inquiry have sought to disentangle identity and task information in EEG data as a means to improve cross-subject generalization. For example, the authors in [24] developed an autoencoder to separate subject-related and task-relevant information. The latent embedding was partitioned into two sections, and the identity content was removed from the first section using an adversarial classifier. A similar result was obtained in [25] using a hierarchical variational autoencoder. Meanwhile, [26] employed three parallel encoders along with an adversarially learned generator/discriminator pair to isolate task-specific, subject-specific, and noise components as a means to improve generalization in motor imagery classification.

In addition to EEG-specific methods, several modality-agnostic strategies have been introduced for general-purpose anonymization. In [27], the authors protected sensitive attributes in biometric data by averaging each sample with a curated set of other examples. Similarly, [28] introduced a method for detecting and obscuring task-irrelevant features using a learnable noise distribution. The authors in [29] extend k-anonymity, a popular method for de-identifying database elements, to time-series data in which specific patterns must be preserved. Their approach, known as (k, P)-anonymity, ensured that at least P dataset elements shared the same temporal pattern, thereby mitigating individual-level identification. Finally, [30] applies a game theoretic model to determine the optimal number of quasi-identifiers to release under privacy and performance constraints.

Prior works in signal-based EEG anonymization such as [15], [16] lack a strong theoretical motivation for their approaches, and hence are difficult to interpret. Additionally, their evaluations are limited to a single attack scenario, in which the attacker trains directly on the protected EEG data. To address these gaps, we devise SAFE based on an interpretable information theoretic model, and evaluate it against three clearly defined privacy attacks that might be launched against an EEG dataset.

III. Methods

A. Framework

Let $\mathcal{D}$ represent a labeled EEG dataset for some classification task $\mathbb{C} : \mathrm{R}^{E\times L} \rightarrow \mathcal{T}$ (e.g., motor imagery, seizure detection) collected from a set of subjects $\mathcal{S}$. Each sample in $\mathcal{D}$ consists of EEG data $x \in \mathrm{R}^{E\times L}$, (where E is the number of channels/electrodes and L is the number of samples), along with class labels $y_{task} \in \mathcal{T}$ for the task $\mathbb{C}$, and subject identity labels $y_{id} \in \mathcal{S}$. Note that each channel reflects the time-series collected by a specific electrode at a specific location on the subject's scalp. We will create and release a modified version of the dataset $\hat{\mathcal{D}} = \mathrm{SAFE}(\mathcal{D})$ instead of the original. We will refer to $\mathcal{D}$ as the 'raw' or 'unprotected' dataset, and $\hat{\mathcal{D}}$ as the 'anonymized' or 'protected' dataset. Our goal is to allow third parties to use $\hat{\mathcal{D}}$ to train on task $\mathbb{C}$ while protecting the identity information of the subjects in $\mathcal{S}$. More specifically, we aim to protect against three distinct attack scenarios, which are defined below. In all cases, the adversary exploits subject labels in one EEG dataset to determine subject identity in another, unlabeled one. This is similar to linkage attacks [31], [32] which also combine multiple information streams to determine an individual's identity. However, our threat models are specified to the unique challenges of EEG data, namely the need to, at times, explicitly release subject labels.

1) Unmasking Attack: In an unmasking attack, it is assumed that the adversary possesses a pre-trained subject-identity classifier (learned using some auxiliary dataset collected from subjects $\mathcal{S}'$). Using this model, the adversary attempts to identify ('unmask') subjects in the protected dataset $\hat{\mathcal{D}}$. Note that under this framework, subject identity labels y_{id} are assumed to **not** be included in $\hat{\mathcal{D}}$, since otherwise the adversary's task would be trivial. As a worst-case scenario, we further assume that the classifier's training data contains the same set of subjects as the anonymized dataset (i.e., $\mathcal{S} = \mathcal{S}'$)[1]. This situation could occur, for example, if the adversary gained access to the subject labels for a subset of $\mathcal{D}$ as the result of a data breach.

2) Inference Attack: The inference attack assumes that subject labels y_{id} *are* included with the released dataset $\hat{\mathcal{D}}$, as is sometimes required when measuring cross-subject generalization is of high importance. The attacker exploits these labels by training a subject classifier on $\hat{\mathcal{D}}$, which they apply to identify subjects within a separate set of unlabeled EEG signals $\tilde{\mathcal{D}}$. For example, $\tilde{\mathcal{D}}$ might represent anonymized medical data that the adversary has procured access to. Once again, we assume a worst-case scenario in which both $\hat{\mathcal{D}}$ and $\tilde{\mathcal{D}}$ are collected from the same set of subjects ($\mathcal{S} = \tilde{\mathcal{S}}$).

3) Open Eval: In addition to protecting subject privacy, the released data must also retain utility for the intended task $\mathbb{C}$. More precisely, a model trained on task $\mathbb{C}$ using $\hat{\mathcal{D}}$ should still be able to make accurate predictions on unseen samples, and models trained on $\mathbb{C}$ using a different dataset should ideally still perform well when evaluated on $\hat{\mathcal{D}}$. Both the inference and unmasking threat models implicitly assume that

[1]In a more realistic scenario, S and S' might have only partial overlap, which would likely degrade the attacker's efficacy.

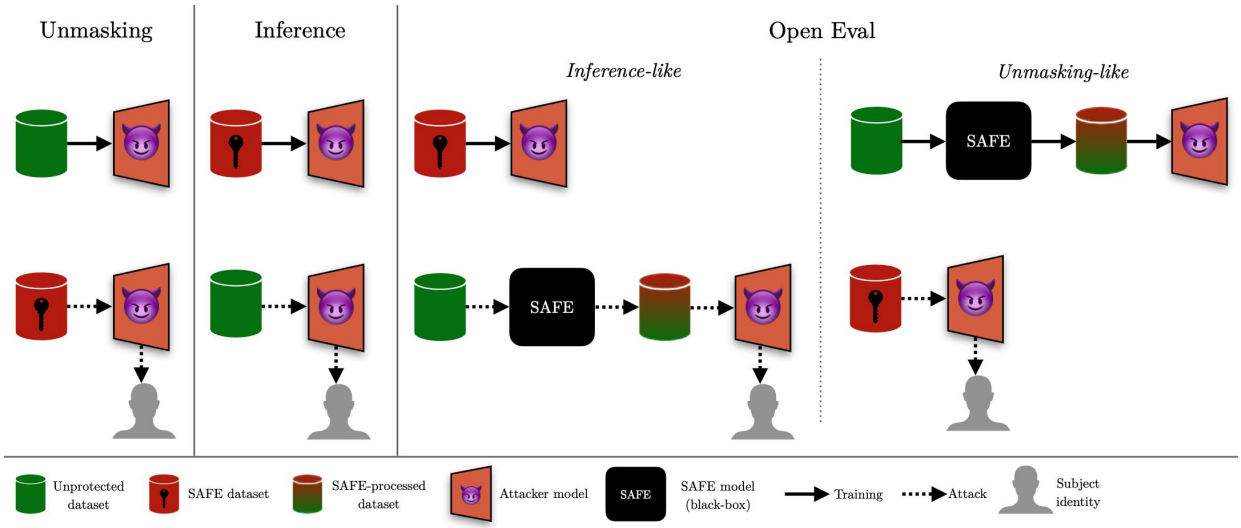

Fig. 1. Overview of the proposed threat models.

the SAFE system is kept fully private, and hence cannot be accessed by adversarial or benign users. While this provides the greatest level of security, it may also inhibit on-task performance. Benign users are not able to apply SAFE to new data when running inference (or apply it to other datasets before training), which creates a distributional shift between training and evaluation. To address this, we also consider a third scenario, which we term open eval, in which the SAFE anonymization model is publicly accessible. It can therefore be utilized by both benign and adversarial users to anonymize additional EEG data as needed. We consider a scenario in which this access is black-box (e.g., via a web API), meaning that a user can query the SAFE model but not access its inner workings. Open eval provides optimal conditions for task utility at the cost of potentially higher vulnerability to privacy attacks. Note that in this situation, there is no practical difference between an inference-like (training on $\hat{\mathcal{D}}$ and running inference on a separate dataset) and an unmasking-like attack (training on a separate dataset and running inference on $\hat{\mathcal{D}}$), since in both cases, the adversary is training and testing on anonymized data. Hence, we use a single experimental paradigm to simulate both such attackers.

B. Anonymization

As before, let the dataset of interest $\mathcal{D} = \{x^{(j)}, y_{task}^{(j)}, y_{id}^{(j)}\}_{j=1}^{N}$ consist of EEG samples $x \in \mathrm{R}^{E\times L}$ with labels $y_{task} \in \mathcal{T}$ and $y_{id} \in \mathcal{S}$, all sampled from some underlying joint distribution $P(X, Y_{task}, Y_{id})$. Rather than directly modifying the raw EEG data itself, we will perform our anonymization in an abstract feature space as defined by a learned mapping $f_{enc} : \mathrm{R}^{E\times L} \to \mathrm{R}^{N_f}$ that decomposes x into a set of features $\mathrm{z} = \{\mathrm{z}_i\}_{i=1}^{N_f}$. This feature-based approach is similar to the one presented in [28]. We will also utilize an (approximate) inverse mapping $f_{dec} : \mathrm{R}^{N_f} \to \mathrm{R}^{C\times T}$ in order to regenerate an EEG-domain sample once the anonymization has been performed (i.e., $f_{dec} \approx f_{enc}^{-1}$). We will denote this reconstruction by $\hat{x} = f_{dec}(z) = f_{dec}(f_{enc}(x)) \approx x$.

The feature set Z represents an abstract latent encoding of the information contained in EEG data X.[2] Consider $\gamma^{(Y_*)}(Z_i) = I(Z_i; Y_*)$, the mutual information [33] between one of the features Z_i and a given label $Y_* \in \{Y_{task}, Y_{id}\}$. If we make the simplifying assumption that the Z_i are mutually independent conditioned on Y_* (meaning essentially that the feature map f_{enc} is non-redundant[3]) we can define

$$\gamma^{(Y_*)}(Z) = I(Z; Y_*) = \sum_{i=1}^{N_f} \gamma^{(Y_*)}(Z_i) = \sum_{i=1}^{N_f} I(Z_i; Y_*) \quad (1)$$

which extends this mutual information measure to the entire feature set Z. Intuitively, $\gamma^{(Y_*)}(Z)$ provides a measure of the utility of Z for predicting label Y_*. Using this formulation, our anonymization objective is to construct a feature set Z that minimizes $\gamma^{(Y_{id})}(Z)$ (to protect the subjects' identity) while simultaneously preserving $\gamma^{(Y_{task})}(Z)$ (so that the dataset can still be employed for its intended task $\mathbb{C}$). While it is possible to achieve this through careful design of the encoder $f_{enc}(\cdot)$, in our implementation, we enhance the anonymization performance by applying a binary mask $M \in \{0, 1\}^{N_f} = \mathcal{M}(X)$ to the feature set Z, giving protected features $\hat{Z} = M \cdot Z$ (where the multiplication is element-wise). Intuitively, $\mathcal{M}(X)$ should be 0 for those features Z_i with $\gamma^{(Y_{id})}(Z_i) \gg \gamma^{(Y_{task})}(Z_i)$ and 1 for those with $\gamma^{(Y_{id})}(Z_i) \ll \gamma^{(Y_{task})}(Z_i)$. After masking, we can then apply the decoder model f_{dec} to map $\hat{Z}$ back to the original domain, generating the anonymized EEG $\hat{X}$.

To make this notion more precise, consider a classifier $h(\cdot)$ that uses the protected signal $\hat{X}$ to predict Y_* (i.e., $\hat{Y}_* = h(\hat{X})$). By invoking Fano's inequality [34] (with the assumption that the prior $P(Y_*)$ is approximately uniform),

[2]We follow the convention of using upper case for random variables and lower case for specific realizations.

[3]For practical encoders learned empirically on a finite dataset (e.g., using a DNN) this independence will likely be only approximately true. However, if N_f is sufficiently small compared with the dimensionality of the raw EEG signal, the model will naturally minimize redundancy to maximize its representational capacity.

we can bound the model's performance by

$$\begin{aligned} P(\hat{Y}_* \neq Y_*) &\geq 1 - \frac{I(\hat{Y}_*; Y_*) + \log 2}{H(Y_*)} \\ &\geq 1 - \frac{I(\hat{Z}; Y_*) + \log 2}{H(Y_*)} \\ &= 1 - \frac{\gamma^{(Y)}(\hat{Z}) + \log 2}{H(Y_*)} \end{aligned} \tag{2}$$

where $H(\cdot)$ is the Shannon entropy. The second bound follows from the data processing inequality and the fact that $Y_* - X - Z - \hat{Z} - \hat{X} - \hat{Y}$ is a Markov chain. In this manner, the utility $\gamma^{(Y_*)}(\hat{Z}) \leq \gamma^{(Y_*)}(Z)$ of the (masked) features $\hat{Z}$ determines a lower bound on the classification accuracy that $h(\cdot)$ can achieve. By properly optimizing $\mathcal{M}$, f_{enc}, and f_{dec} to minimize $\gamma^{(Y_{id})}(\hat{Z})$ while preserving $\gamma^{(Y_{task})}(\hat{Z})$, we can therefore enforce anonymization while simultaneously ensuring that the dataset remains useful for the intended task $\mathbb{C}$. We can now formulate our anonymization objective as follows:

$$\begin{aligned} \text{SAFE}(\mathcal{D}) = \Big\{\Big(& \\ & f^*_{dec}\big(\mathcal{M}^*(x^{(j)})\big) \cdot f^*_{enc}(x^{(j)}), \\ & y^{(j)}_{task},\ y^{(j)}_{id}\Big)\Big\}_{j=1}^{N} \\ \text{where} \quad f^*_{enc}, f^*_{dec}, \mathcal{M}^* = \arg & \max_{f_{enc}, f_{dec}, \mathcal{M}} \\ & \mathbb{E}\Big[\gamma^{(Y_{task})}(\hat{Z}) - \lambda\gamma^{(Y_{id})}(\hat{Z})\Big] \\ \text{s.t.} \quad \hat{Z} &= \mathcal{M}(X) \cdot f_{enc}(X), \\ f_{dec} &\approx f^{-1}_{enc} \end{aligned} \tag{3}$$

Where $\lambda \in R^+$ is a hyperparameter controlling the tradeoff between anonymization and task performance.

We now consider this framework in light of the threat models defined in the previous section. In an unmasking attack, the adversary's pre-trained classifier will have its performance limited when evaluated on $\hat{\mathcal{D}}$, as described by the bounds above. They will therefore be unable to accurately identify the subject in the protected data. In an inference attack, the adversary will attempt to train his or her model on $\hat{\mathcal{D}}$. However, once again, the bounds in 2 will prevent this model from being effectively optimized. One subtle point in this scenario is that the accuracy bounds apply only in expectation over the entire data distribution. Thus, the attacker *may* still achieve robust performance on the finite dataset $\hat{\mathcal{D}}$ due to, for example, overfitting or memorization. However, this model will fail to generalize when it is applied to identify subjects in a different dataset. Finally, in the open eval scenario, the attacker can pass their training and or target data through the SAFE anonymization system. While this process will reduce distributional shifts between their training and inference data, the suppression of identity features will continue to constrain their model's performance, just as in the inference and unmasking threat models.

IV. Method

A. Model

Inspired by the approaches in [15] and [24], we used an autoencoder-style model to parameterize the feature decomposition and reconstruction functions f_{enc} and f_{dec}. Meanwhile, a separate model g_{mask} was utilized to predict the optimal latent space mask conditioned on the raw EEG data as input (thereby parameterizing the masking function $\mathcal{M}$). Rather than attempting to directly learn feature utility values $\gamma^{(Y_*)}(Z_i)$, we instead optimize 3 implicitly using adversarial learning. Two classifiers, $h_{task}(\cdot)$ and $h_{id}(\cdot)$, for task labels y_{task} and subject labels y_{id}, respectively, were trained on the reconstructed data $\hat{x}$ generated from the output of f_{dec}. Backpropagating these models' losses to the autoencoder and masking model pushed them to remove identity features while retaining task information. All models f_{enc}, f_{dec}, $h_{task}(\cdot)$, $h_{id}(\cdot)$, and g_{mask} were jointly optimized. The autoencoder and masking model were trained using a weighted combination of the task model loss, the negative subject model loss, an MSE reconstruction penalty, and a KL-divergence consistency loss using the task model's logits. The MSE and KL losses were employed to minimize distribution shift in the anonymized data, enforcing that $f_{dec} \approx f^{-1}_{enc}$. We provide further details on each of these components in the following sections.

1) Autoencoder: The autoencoder $f = f_{dec} \circ f_{enc}$ consists of an encoder module $f_{enc}(\cdot)$ which takes in raw EEG signal x and produces latent feature vector z along with a decoder $f_{dec}(\cdot)$ that reconstructs the anonymized EEG $\hat{x}$. We based the architecture of our autoencoder on the one presented in [35]. The encoder portion uses EEGNet [36], which consists of time-wise, depth-wise, and separable convolution stages. Both the depth-wise and separable convolutions are followed by average pooling to reduce the dimension of the internal feature maps. For the decoder, we use a sequence of transposed convolutions to invert the down-sampling in the decoder. A final upsampling layer is used to correct any small shape mismatches between the original and reconstructed EEG.

2) Masking Model: The masking model $g_{mask}(\cdot)$ consists of two parallel EEGNet architectures that predict parameters $\log \alpha$ and β conditioned on the raw EEG signal x as input. These parameters are then used to sample a stochastic multiplicative mask $m^{(x)}$ that is applied element-wise to the latent representation z. The use of a semi-randomized masking function (as opposed to a deterministic one) reduces the potential for an adversary to discover or invert the inner workings of the SAFE system. In order to make the process of sampling this mask differentiable, we use an approach based on the hard concrete distribution from [37]. For each latent feature z_i, we sample an independent uniform random variable $u_i \sim \text{Unif}[0, 1]$. Next, we apply an element-wise sigmoid function

$$s = \sigma\left(\frac{\log u - \log(1-u) + \log \alpha}{\beta}\right)$$

where the parameters $\log \alpha$ and β are the ones predicted by the masking model (and thus are functions of the EEG signal x). Next, s is shifted and scaled using fixed hyperparameters ω, and ζ to give $s' = s(\zeta - \omega) + \omega$. Finally, the mask m is computed as $m = \min(1, \max(0, s'))$. Note that, to aid in

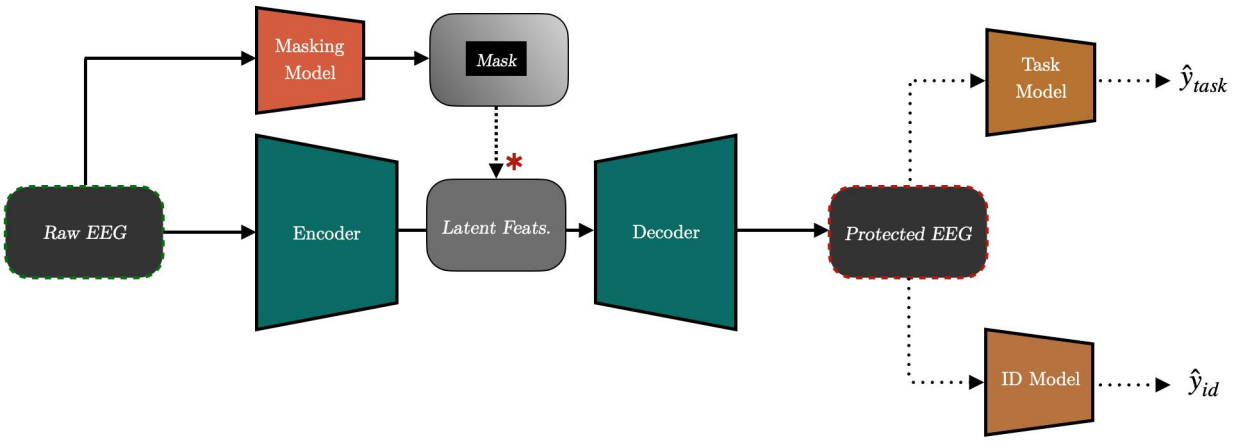

Fig. 2. Schematic diagram of the SAFE model.

learning, we have slightly relaxed our theoretical formulation by allowing mask elements that are not strictly 0 or 1. However, we observed in practice that most mask elements were indeed close to these extreme values, and that imposing hard binary masks when anonymizing had minimal impact on the performance.

3) Task and Subject Models: To train the anonymization system, we must enforce the suppression of identity features while preserving features that are important to the task, as expressed in equation 3. To achieve this, we also train task and identity classifiers $h_{task}(\hat{x})$ and $h_{id}(\hat{x})$ in parallel with the masking model and autoencoder. Note that both classifiers are trained on the *reconstructed* EEG data. The loss from the task model is back-propagated through the autoencoder/masking model backbone such that they are encouraged to retain task-relevant features. Meanwhile, the loss from the subject classifier is used adversarially to penalize the propagation of subject identity. Both classifiers use an EEGNet architecture.

B. Training

For each training batch, the signal x is passed through the autoencoder and masking backbone to generate protected reconstruction $\hat{x}$, which is in turn passed through the task and identity classifiers $h_{task}(\hat{x})$ and $h_{id}(\hat{x})$. We then compute the anonymization loss:

$$\mathcal{L}_{anon} = \theta_{task}\mathcal{L}_{task} + \theta_{mse}\mathcal{L}_{mse} + \theta_{kl}\mathcal{L}_{kl} - \theta_{id}\mathcal{L}_{id}.$$

Here $\mathcal{L}_{task}$ and $\mathcal{L}_{id}$ are the cross-entropy classification losses for the task and identity models, and $\mathcal{L}_{mse}$ is the MSE reconstruction loss between the original and anonymized data. Note that we use the *negative* of $\mathcal{L}_{id}$ since the subject model is adversarial. To further minimize potential distributional shifts in the task-relevant features between the original and anonymized data, we also introduce $\mathcal{L}_{kl}$, a KL divergence penalty computed between the task-model's logits for x and $\hat{x}$: $\mathcal{L}_{kl} = \mathrm{KL}(\mathrm{h_{task}(x)}; \mathrm{h_{task}(\hat{x})})$. The θ_* represent weighting terms that define the relative contribution of each loss component. These are set to maximize performance and also define the tradeoff between privacy and task accuracy. We backpropagate $\mathcal{L}_{anon}$ to update the weights of the autoencoder and masking model, with the task and subject models frozen. Next, we update both $h_{task}(\cdot)$ and $h_{id}(\cdot)$ using their respective cross-entropy loss functions $\mathcal{L}_{task}$ and $\mathcal{L}_{id}$. In practice, we often found it beneficial to perform multiple updates to the adversarial subject model on each batch so that it remained competitive with the anonymization.

V. Experiments

A. Datasets

We evaluate our method using 4 EEG classification datasets representing a variety of tasks. Each dataset possesses a trial-based structure, in which classification labels are predicted from relatively short EEG segments (approximately $300-1500$ samples).

1) BCI IV Dataset 2a: The BCI IV competition introduced several datasets to support the development of brain-computer interfaces (BCIs). The 2a subset [38] contains EEG from nine subjects performing a motor imagery task in which they imagined moving either their left hand, right hand, feet, or tongue. The objective is to distinguish between these four classes using the EEG recordings. The EEG was captured using 22 electrodes in a 10-20 layout [39] with a sampling rate of 250 Hz. In addition to the high and low pass filtering pre-applied to the released data, we also added a 60 Hz notch filter to remove power line noise. We then performed channel-wise z-score normalization and ICA for blink removal. The data was epoched by extracting the time window from -0.2s to 4.0s around each cue presentation (1051 samples). We use only the labeled training sessions for our experiments, comprising a total of 288 trials from each participant.

2) GigaDB Dataset: In [40] Cho *et. al.* presented a motor imagery dataset from 52 individuals, in which subjects imagined moving either their left or right hand. A total of 100 to 120 trials with each hand were collected per participant. The target task is to classify which hand the subject is imagining moving. The EEG was recorded with 64 electrodes following the 10-10 layout [41] and was originally sampled at 512 Hz. We applied high and low-pass filtering (at 0.5 Hz and 100

TABLE I
BALANCED ACCURACY PERFORMANCE (↑) OF TASK MODELS UNDER DIFFERENT TRAINING AND EVALUATION SCENARIOS (ORIG = ORIGINAL AND ANON = SAFE ANONYMIZED).

Dataset	Random Chance	Orig-train / Orig-eval	Orig-train / Anon-eval	Anon-train / Orig-eval	Anon-train / Anon-eval
MMI	0.5	0.828 ± 0.014	0.753 ± 0.054	0.711 ± 0.027	0.810 ± 0.018
BCI IV 2a	0.25	0.494 ± 0.059	0.395 ± 0.059	0.309 ± 0.030	0.467 ± 0.074
SMNI-stim	0.33	0.677 ± 0.017	0.451 ± 0.034	0.395 ± 0.030	0.571 ± 0.021
SMNI-group	0.5	0.677 ± 0.053	0.535 ± 0.051	0.566 ± 0.037	0.646 ± 0.046
GigaDB	0.5	0.669 ± 0.042	0.591 ± 0.053	0.578 ± 0.053	0.642 ± 0.047

TABLE II
BALANCED ACCURACY PERFORMANCE (↓) OF ADVERSARIAL IDENTIFICATION MODEL UNDER DIFFERENT THREAT MODELS (ORIG = ORIGINAL AND ANON = SAFE ANONYMIZED).

Dataset	Random Chance	Orig-train / Orig-eval	Orig-train / Anon-eval	Anon-train / Orig-eval	Anon-train / Anon-eval
Adversary mode		*Undefended*	*Unmasking*	*Inference*	*Open Eval*
MMI	0.010	0.956 ± 0.011	0.017 ± 0.006	0.017 ± 0.006	0.122 ± 0.015
BCI IV 2a	0.111	0.991 ± 0.008	0.186 ± 0.043	0.161 ± 0.032	0.686 ± 0.051
SMNI-stim	0.008	0.878 ± 0.035	0.011 ± 0.002	0.020 ± 0.006	0.119 ± 0.011
SMNI-group	0.008	0.876 ± 0.046	0.011 ± 0.003	0.020 ± 0.005	0.146 ± 0.015
GigaDB	0.019	0.997 ± 0.002	0.033 ± 0.010	0.033 ± 0.008	0.645 ± 0.043

Hz), notch filtering at 60 Hz, z-score normalization, and ICA blink removal. Each example was 1639 samples in length.

3) MMI Dataset: The Motor Movement/Imagery (MMI) dataset [42] consists of EEG recordings from a combination of real and imagined movements. For our study, we used the set of trials in which the participants were instructed to open and close either their left or right hand (real movements), with the target objective of determining which hand was being used. A total of three two-minute runs of this protocol were collected from each subject. Although the original dataset consisted of 109 subjects, we used only 105, dropping 4 participants with data issues. The EEG was recorded at 160 Hz using 64 electrodes in a 10-10 layout. We applied high and low-pass filtering (at 0.5 Hz and 50 Hz) and notch filtering at 60 Hz for preprocessing. We also employed channel-wise z-score normalization and ICA for blink removal. Each trial was 480 samples in length.

4) SMNI Dataset: The SMNI Dataset [43] contains EEG recordings from both alcoholic and control participants. Each subject was exposed to either a single object, two matched objects, or two non-matched objects. We used data from a total of 121 subjects (dropping one subject with missing data), with 120 stimulus trials for each subject. We tested both group classification (i.e., distinguishing between control and alcoholic individuals) and stimulus classification tasks. The EEG was sampled at 256 Hz, using 64 electrodes, with one second of data recorded for each trial. We mapped the electrodes into a standard 10-20 montage with 50 channels in order to interpolate missing or bad channels. Channel-wise z-score normalization was also applied for each trial.

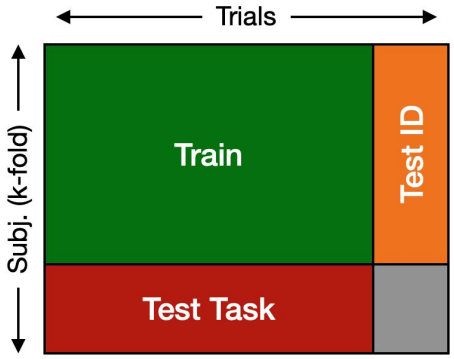

Fig. 3. Per-fold splitting strategy used for training and evaluating the SAFE model. Folds are defined across subjects (subj.).

B. Evaluation Method

We used 5-fold cross-validation to evaluate the performance of the proposed method. The folds were stratified across subjects, such that for each fold, $\sim$ 20% of the subjects were set aside for testing the performance of the task model (the remaining 80% were used for training). We next further subdivided the *trials* from the training subjects to hold out 20% for testing the identity classifier, leaving the remainder for training. See Figure 3 for a visual depiction of the splitting within a given fold. This process was necessary so that task models could be evaluated on cross-subject performance, while identity models could be evaluated on different trials from the same set of subjects. For each fold, the training split was used to learn the SAFE anonymization model, which was subsequently applied to anonymize both the training and testing splits.

To rigorously evaluate utility and anonymization for the threat models defined in section III, we trained and evaluated task and subject classifiers in four different settings:

1) **Raw training/Raw testing:** This acts as a baseline for both task performance and susceptibility to privacy attacks in the case where no anonymization is applied.
2) **Raw training/Anon. testing:** For the identity model, this simulates an unmasking attack, where the attacker pretrains a classifier on unprotected data and then applies it to the SAFE-protected dataset. For the task model, it emulates a user that applies a pre-trained model (learned using unprotected EEG) to the anonymized data.
3) **Anon. training/Raw testing:** For the identity model, this simulates an inference attacker who tries to use the protected dataset to train a classifier, which they then can apply to identify subjects in a separate unprotected dataset. For the task model, it simulated a user training on the SAFE data and then evaluating using unprotected EEG.
4) **Anon. training/Anon testing:** This simulates the open eval scenario, where both benign and adversarial users can freely anonymize additional data using the SAFE system. Recall that in this case, the inference-like and unmasking-like attacks are symmetric because both use protected data for training and adversarial evaluation.

Note that both model types were trained using the training split (described above), but evaluated using either the cross-subject or cross-trial test subset as appropriate. In an ideal case, an additional set of training samples (separate from the one used to learn the SAFE model) would be employed to simulate the raw pretraining data in setting 2. However, the limited size of the datasets utilized made this infeasible. We nonetheless believe that our setup provides a reasonable measure of the SAFE method's performance, since biases introduced by optimizing the SAFE system on these same data should be relatively minimal, given that performance was still evaluated on a separate test set that was withheld from all models.

We used balanced accuracy (BA) to measure both task performance and attacker success. Balanced accuracy is more robust to class imbalances and, hence, provides a more calibrated performance measure. We performed our main evaluations using 4 different random seeds, and report both the average performance (across all seeds and folds) as well as the standard deviation. Initial hyperparameter tuning was performed using a different random seed from the ones used for final evaluation. We note that our setup differs from conventional model training in that the objective is to develop a SAFE model that provides maximal protection for a *specific* dataset, rather than one that generalizes to unseen data.

C. Implementation

We implemented the SAFE method using PyTorch [44] and PyTorch Lightning [45]. For the BCI dataset, training the autoencoder on 2 NIVIDIA V100 GPUs took approximately 8.5 minutes, while anonymization took less than 15 seconds. We use the TorchEEG [46] implementations of the EEGNet and FBCNet [47] models, and the official repository for the TSCeption model [48]. The files for the GigaDB dataset were sourced from [49].

VI. Results and Discussion

A. Task Performance

Table I shows the performance of the task models across the different training and evaluation scenarios described above. The first column (training and evaluating on the original data) serves as a baseline measure of the original dataset's utility for a benign actor, who uses the data for its intended purpose. This baseline performance varies substantially across the datasets, indicating that some tasks and corpora (e.g., GigaDB) are quite difficult, while others (e.g., MMI) are easier to model.

The remaining three columns reflect the performance cost incurred as a result of SAFE anonymization. Column two (training on original data and testing on the anonymized data) simulates a user who runs inference on the anonymized data using a model trained on unaltered EEG signals. This domain mismatch causes a performance drop across all datasets of roughly $8\% - 12\%$, with both SMNI tasks exhibiting the greatest degradation. Meanwhile, column three captures the reverse scenario, in which a model trained on the anonymized data is applied to a set of unmodified samples. Once again, task accuracy is decreased compared to the baseline. The magnitude of this degradation is slightly higher than in the original-to-anon configuration ($\sim 10\% - 18\%$), perhaps because the models tend to overfit to specific artifacts in the SAFE-protected data.

TABLE III
IDENTIFICATION ACCURACY (↓) UNDER THREAT MODELS WITH/WITHOUT MASKING (LABELED M/NM) ON THE MMI DATASET. RESULTS ARE SHOWN FOR THE UNMASKING (UNM.), INFERENCE (INF.), AND OPEN EVAL (O.E.) THREAT MODELS.

Dataset	Unm.		Inf.		O.E.	
	M	NM	M	NM	M	NM
MMI	0.017	0.017	0.017	0.015	0.122	0.218
BCI	0.186	0.149	0.161	0.180	0.686	0.784
SMNI-s	0.011	0.013	0.020	0.017	0.119	0.201
SMNI-g	0.011	0.014	0.020	0.017	0.146	0.247
GigaDB	0.033	0.030	0.033	0.030	0.645	0.776

The performance gap incurred when evaluating across schemas (i.e., clean to anon or vice versa) indicates that our method is not able to fully preserve task-relevant features from the clean data. However, it may also be the case that there exists a certain degree of entanglement between identity and task features, and that suppressing the former has certain irreconcilable effects on the latter. This hypothesis is supported by the variability in the performance gap across datasets, since different tasks, collection protocols, etc., would likely create different degrees of feature overlap.

The final column in Table I depicts the open eval situation in which the user trains a model on protected data and also applies the same anonymization protocol to the testing data before running inference. The accuracy gap under this

scenario is far smaller (roughly $2\% - 4\%$) than for the cross-schema evaluations. Notably, the SMNI stimulus task exhibits a larger gap than the other datasets. The fact that performance degradation is largely mitigated by applying SAFE to both training and testing data indicates that the task models can adapt to leverage different feature subsets that are better preserved in the anonymization process, yet still enable robust and generalizable task performance, provided that distribution shift is minimized. In other words, the anonymization tends to *modify* the subset of task-relevant features more so than remove them outright.

B. Identity Protection

Table II shows the performance of the simulated privacy adversary under each of the threat models. The unprotected baseline accuracy (column one) is very high across all datasets, and essentially perfect for BCI and GigaDB. This indicates that the EEG signals contain strong biomarkers of subject identity, validating the findings in prior work (e.g., [7]) and motivating the need for anonymization methods. Note that this scenario provides a baseline for both inference and unmasking threat models, since they are fully symmetric in the absence of defensive measures. Columns two and three show the defended performance of the unmasking and inference attackers, respectively. In both cases, the attacker's success is decreased drastically (to near random chance) for all datasets, indicating that SAFE provides a robust defense that largely thwarts the adversary's efforts. No consistent difference in adversarial performance is observed between the two threat models.

However, as discussed in the previous section, benign users incur a non-trivial cost in task accuracy when models trained on protected data are evaluated on a clean dataset, and vice versa. Thus, depending on the relative tradeoff between identity suppression and task utility, it may be advantageous to publicize the SAFE model. This results in the open eval scenario, which is shown in the final column of Table II. While anonymization performance is degraded over the other threat models (in which the SAFE model is kept secret), our method still provides fairly robust privacy protection, especially for the MMI and SMNI datasets. The fact that BCI and GigaDB are less well protected may indicate that these datasets contain stronger identity markers, perhaps exacerbated by differences in recording conditions between subjects. This is consistent with the fact that both these corpora also admit near-perfect identification performance when left unprotected.

Overall, SAFE exhibits strong identity protection across the different datasets, even against the most challenging threat model (open eval). In all cases, SAFE induces a statistically significant decrease in identification accuracy ($p \ll 0.01$) as measured by a Wilcoxon signed-rank test across seeds and folds. While on-task performance is degraded when evaluating across schemas (e.g., clean to protected), this degradation may be acceptable in certain highly sensitive applications in which maximal anonymization is imperative. Furthermore, in cases such as federated learning, where model training is performed by a (presumably trustworthy) centralized actor, the performance benefits of the open eval scenario can be realized without needing to publicize the anonymization system.

C. Impact of Latent Space Masking

In the SAFE model, both the latent-space masking and the autoencoder structure contribute to EEG anonymization via the removal of identity features. To assess the relative role that each plays, we ran a series of ablations (with the same 5-fold structure as the main experiments) in which we trained SAFE models without the masking mechanism. The results are shown in Table III. Surprisingly, for both the unmasking (Unm.) and inference (Inf.) threat models, the differences between masked and unmasked performance are minimal, indicating that the autoencoder itself is sufficient to guard against these attackers. However, in the more challenging open eval (O.E.) scenario, the mask substantially improves the defense's efficacy, decreasing the identification accuracy by roughly 10% across all datasets. This indicates that explicit masking of latent features is critical to defending against more robust attackers. We note that this represents an advance over prior anonymization methods (e.g., [15]) that have relied solely on the implicit information removal induced by the autoencoder's compressed representation.

D. Sensitivity to Attacker Model Architecture

TABLE IV
IDENTIFICATION ACCURACY (↓) WHEN THE ATTACKER USES A DIFFERENT MODEL THAN THE SAFE SYSTEM.

Model	Unprotected	Unmasking	Inf.	Open Inf.
EEGNet	*0.956*	*0.017*	*0.017*	*0.122*
TSception	0.845	0.014	0.019	0.050
FBCNet	0.823	0.014	0.013	0.069

SAFE uses an adversarial subject-prediction model based on EEGNet as a core part of its anonymization procedure. However, our defense must remain robust even against an adversary who selects a different architecture to use in their attack. To verify this, we ran an ablation in which we used the TSception [48] and FBCNet [47] models to test the attacker's performance in subject identification. We used the MMI dataset, with the same setup as the main experiments. Results are shown in Table IV, with the top row representing the simulated EEGNet adversary (identical to Table II). The second and third rows, which report the results for TSception and FBCNet, respectively, demonstrate that SAFE is still effective in protecting identity information from an unmatched adversary. This indicates that SAFE has suppressed fundamental identity features that are model-invariant, which validates the information-theoretic nature of our approach. It further suggests that our method does not substantially overfit to the specific adversary used during training.

VII. CONCLUSION

In this paper, we have presented a novel system for anonymizing an EEG dataset, say prior to public release. Our

approach, SAFE, employs an autoencoder in concert with latent space feature masking to selectively remove identity information while protecting task-relevant features. In addition to establishing a theoretical basis for our method, we clearly define a set of threat models by which an attacker might maliciously exploit identity information in the published data. Using four popular EEG datasets, we simulate these attackers empirically and report both adversarial and on-task performance. Our approach markedly decreased the identification abilities of both the unmasking and inference attackers, at a moderate performance cost for benign dataset users. In the open eval case, where the anonymization system is publicly accessible, the reduction in task performance is substantially reduced, and anonymization protections remain relatively robust. Taken collectively, our results represent a promising new approach to privacy-preserving deep learning that mitigates the unauthorized exploitation of identity information in published EEG datasets. Providing these protections will be crucial to safely realizing the full benefits of contemporary AI methods for neurological insight. However, we do not evaluate our method against a white-box adversary that might attempt to directly circumvent the anonymization model itself, and only consider trial-based EEG classification datasets. Future work should consider how our approach might be modified to accommodate a broader range of EEG tasks, particularly those that involve longer data segments.

VIII. Acknowledgment

Many thanks to Aditya Kommineni for providing feedback on the initial idea, as well as code for the MMI dataset. Thanks to Kleanthis Avramidis for his helpful feedback.

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the ARTS Program under contract D2023-2308110001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

IX. Appendix I

A. Derivation of Mutual Information form of Fano's Inequality (based on [50])

We start with the standard form of Fano's inequality with $P_e = P(Y' \neq Y)$

$$H(Y|Y') \leq H_b(P_e) + P_e \log(|\mathcal{Y}| - 1)$$

where $H_b(p) = -p \log p - (1-p)\log(1-p) \leq \log 2$ is the binary entropy function and random variable Y has discrete support set $\mathcal{Y}$. We therefore have:

$$H(Y|Y') - H(Y) \leq H_b(P_e) + P_e \log(|\mathcal{Y}| - 1) - H(Y)$$

$$-I(Y;Y') \leq H_b(P_e) + P_e \log(|\mathcal{Y}| - 1) - H(Y)$$

$$H(Y) - I(Y;Y') \leq \log 2 + P_e \log(|\mathcal{Y}|)$$

If we assume the prior on Y is uniform over $\mathcal{Y}$, then $H(Y) = \log|\mathcal{Y}|$ and so

$$\frac{\log|\mathcal{Y}| - I(Y;Y') - \log 2}{\log(|\mathcal{Y}|)} \leq P_e$$

or

$$1 - \frac{I(Y;Y') + \log 2}{H(Y)} \leq P_e$$

[1] S. Gannouni, A. Aledaily, K. Belwafi, H. Alshazly, and E. Ben Hamida, "Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification," *Scientific Reports*, vol. 11, no. 1, p. 7071, 2021.

[2] M. Jafari, A. Shoeibi, M. Khodatars, S. Bagherzadeh, A. Shalbaf, D. L. García, J. M. Gorriz, and U. R. Acharya, "Emotion recognition in eeg signals using deep learning methods: A review," *Computers in Biology and Medicine*, vol. 165, p. 107450, 2023.

[3] G. Yogarajan, N. Alsubaie, G. Rajasekaran, H. Alshazly, and E. Ben Hamida, "Eeg-based epileptic seizure detection using binary dragonfly algorithm and deep neural network," *Scientific Reports*, vol. 13, no. 1, p. 17710, 2023.

[4] X. Wang, Y. Wang, D. Liu, X. Ma, and C. Yin, "Automated recognition of epilepsy from eeg signals using a combining space–time algorithm of cnn-lstm," *Scientific Reports*, vol. 13, no. 1, p. 14876, 2023.

[5] N. Sharma, A. Upadhyay, M. Sharma, Y. Liu, and M. Ahmadi, "Deep temporal networks for eeg-based motor imagery recognition," *Scientific Reports*, vol. 13, no. 1, p. 18813, 2023.

[6] N. Tibrewal, N. Leeuwis, and M. Alimardani, "Classification of motor imagery eeg using deep learning increases performance in inefficient bci users," *PLOS ONE*, vol. 17, pp. 1–18, 07 2022.

[7] Q. Gui, M. V. Ruiz-Blondet, S. Laszlo, and Z. Jin, "A survey on brain biometrics," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–38, 2019.

[8] S. Bak and J. Jeong, "User biometric identification methodology via eeg-based motor imagery signals," *IEEE Access*, vol. 11, pp. 41303–41314, 2023.

[9] A. B. Tatar, "Biometric identification system using eeg signals," *Neural Computing and Applications*, vol. 35, pp. 1009–1023, 2023.

[10] M. J. Rivera, M. A. Teruel, A. Maté, and J. Trujillo, "Diagnosis and prognosis of mental disorders by means of EEG and deep learning: a systematic mapping study," *Artificial Intelligence Review*, vol. 55, pp. 1209–1251, 2022.

[11] H. Uyanik, A. Sengur, M. Salvi, R.-S. Tan, J. H. Tan, and U. R. Acharya, "Automated detection of neurological and mental health disorders using eeg signals and artificial intelligence: A systematic review," *WIREs Data Mining and Knowledge Discovery*, vol. 15, no. 1, p. e70002, 2025. e70002 DMKD-00820.R1.

[12] V. Joshi and N. Nanavati, "A review of eeg signal analysis for diagnosis of neurological disorders using machine learning," *Journal of Biomedical Photonics & Engineering*, vol. 7, no. 4, p. 040201, 2021.

[13] C. Micanovic and S. Pal, "The diagnostic utility of eeg in early-onset dementia: a systematic review of the literature with narrative analysis," *Journal of Neural Transmission (Vienna)*, vol. 121, pp. 59–69, Jan. 2014. Epub 2013 Jul 31.

[14] Y. Liu, Y. Chen, G. Fraga-González, V. Szpak, J. Laverman, R. W. Wiers, and K. R. Ridderinkhof, "Resting-state eeg, substance use and abstinence after chronic use: A systematic review," *Clinical EEG and Neuroscience*, vol. 53, no. 4, pp. 344–366, 2022. PMID: 35142589.

[15] G. Singh, P. Patel, M. Asaduzzaman, and G. Bajwa, "Selective eeg signal anonymization using multi-objective autoencoders," in *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*, pp. 1–7, 2023.

[16] L. Meng, X. Jiang, J. Huang, W. Li, H. Luo, and D. Wu, "User identity protection in eeg-based brain–computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3576–3586, 2023.

[17] E. Debie, N. Moustafa, and M. T. Whitty, "A privacy-preserving generative adversarial network method for securing eeg brain signals," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020.

[18] H. Liu, Y. Zhang, Y. Li, and X. Kong, "Review on emotion recognition based on electroencephalography," *Frontiers in Computational Neuroscience*, vol. 15, 2021.

[19] S. M. Park, B. Jeong, D. Y. Oh, C.-H. Choi, H. Y. Jung, J.-Y. Lee, D. Lee, and J.-S. Choi, "Identification of major psychiatric disorders from resting-state electroencephalography using a machine learning approach," *Frontiers in Psychiatry*, vol. 12, 2021.

[20] D. Pascual, A. J. Amirshahi, A. Aminifar, D. A. Alonso, P. Ryvlin, and R. Wattenhofer, "Epilepsygan: Synthetic epileptic brain activities with privacy preservation," *IEEE Transactions on Biomedical Engineering*, vol. 68, pp. 2435–2446, 2020.

[21] A. Nolin-Lapalme, R. Avram, and H. Julie, "Privecg: generating private ecg for end-to-end anonymization," in *Proceedings of the 8th Machine Learning for Healthcare Conference* (K. Deshpande, M. Fiterau, S. Joshi, Z. Lipton, R. Ranganath, I. Urteaga, and S. Yeung, eds.), vol. 219 of *Proceedings of Machine Learning Research*, pp. 509–528, PMLR, 11–12 Aug 2023.

[22] A. Agarwal, R. Dowsley, N. D. McKinney, D. Wu, C.-T. Lin, M. De Cock, and A. C. A. Nascimento, "Protecting privacy of users in brain-computer interface applications," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 8, pp. 1546–1555, 2019.

[23] V. Robinson and E. B. Varghese, "A novel approach for ensuring the privacy of eeg signals using application-specific feature extraction and aes algorithm," *2016 International Conference on Inventive Computation Technologies (ICICT)*, vol. 2, pp. 1–6, 2016.

[24] M. Han, O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Disentangled adversarial autoencoder for subject-invariant physiological feature extraction," *IEEE signal processing letters*, vol. 27, pp. 1565–1569, 2020.

[25] L. Bollens, T. Francart, and H. V. Hamme, "Learning subject-invariant representations from speech-evoked eeg using variational autoencoders," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1256–1260, 2022.

[26] J. Han, X. Gu, G.-Z. Yang, and B. Lo, "Noise-factorized disentangled representation learning for generalizable motor imagery eeg classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 2, pp. 765–776, 2024.

[27] C.-F. Chen, B. Moriarty, S. Hu, S. Moran, M. Pistoia, V. Piuri, and P. Samarati, "Model-agnostic utility-preserving biometric information anonymization," *International Journal of Information Security*, pp. 1–18, 2024.

[28] F. Mireshghallah, M. Taram, A. Jalali, A. T. Elthakeb, D. M. Tullsen, and H. Esmaeilzadeh, "Not all features are equal: Discovering essential features for preserving prediction privacy," *Proceedings of the Web Conference 2021*, 2021.

[29] L. Shou, X. Shang, K. Chen, G. Chen, and C. Zhang, "Supporting pattern-preserving anonymization for time-series data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 877–892, 2013.

[30] L. Kuang, Y. Zhu, S. Li, X. Yan, H. Yan, and S. Deng, "A privacy protection model of data publication based on game theory," *Security and Communication Networks*, vol. 2018, no. 1, p. 3486529, 2018.

[31] A. Vidanage, T. Ranbaduge, P. Christen, and R. Schnell, "A taxonomy of attacks on privacy-preserving record linkage," *Journal of Privacy and Confidentiality*, vol. 12, Jul. 2022.

[32] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, 2008.

[33] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[34] R. M. Fano, *Transmission of Information: A Statistical Theory of Communications*. MIT Press, 1961.

[35] D. Bethge, P. Hallgarten, T. Grosse-Puppendahl, M. Kari, L. L. Chuang, O. Özdenizci, and A. Schmidt, "Eeg2vec: Learning affective eeg representations via variational autoencoders," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3150–3157, IEEE, 2022.

[36] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[37] C. Louizos, M. Welling, and D. P. Kingma, "Learning sparse neural networks through l_0 regularization," 2018.

[38] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "Bci competition 2008 – graz data set a." `http://www.bbci.de/competition/iv/`, 2008. Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology.

[39] G. H. Klem, H. O. Lüders, H. H. Jasper, and C. Elger, "The ten-twenty electrode system of the international federation. the international federation of clinical neurophysiology," *Electroencephalography and Clinical Neurophysiology. Supplement*, vol. 52, pp. 3–6, 1999.

[40] H. Cho, M. Ahn, S. Ahn, M. Kwon, and S. C. Jun, "EEG datasets for motor imagery brain–computer interface," *GigaScience*, vol. 6, no. 7, 2017.

[41] American Electroencephalographic Society, "Guideline thirteen: guidelines for standard electrode position nomenclature," *Journal of Clinical Neurophysiology*, vol. 11, pp. 111–113, Jan. 1994.

[42] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1034–1043, Jun 2004.

[43] H. Begleiter, "EEG Database." UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C5TS3D.

[44] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[45] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," Mar. 2019.

[46] Z. Zhang, S. hua Zhong, and Y. Liu, "TorchEEGEMO: A deep learning toolbox towards EEG-based emotion recognition," *Expert Systems with Applications*, p. 123550, 2024.

[47] R. Mane, E. Chew, K. Chua, K. K. Ang, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, "Fbcnet: A multi-view convolutional neural network for brain-computer interface," 2021.

[48] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, "Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2238–2250, 2022.

[49] B. Aristimunha, I. Carrara, P. Guetschel, S. Sedlar, P. Rodrigues, J. Sosulski, D. Narayanan, E. Bjareholt, Q. Barthelemy, R. T. Schirrmeister, R. Kobler, E. Kalunga, L. Darmet, C. Gregoire, A. Abdul Hussain, R. Gatti, V. Goncharenko, J. Thielen, T. Moreau, Y. Roy, V. Jayaram, A. Barachant, and S. Chevallier, "Mother of all bci benchmarks," 2025.

[50] J. Scarlett and V. Cevher, "An introductory guide to fano's inequality with applications in statistical estimation," 2019.