Domain-guided Self-supervision of EEG Data Improves Downstream Classification Performance and Generalizability

Neeraj Wagh NWAGH2@ILLINOIS.EDU
Jionghao Wei WEI33@ILLINOIS.EDU
Samarth Rawal SCRAWAL2@ILLINOIS.EDU
Brent Berry BERRY.BRENT@MAYO.EDU
Leland Barnard BARNARD.LELAND@MAYO.EDU
Benjamin Brinkmann BRINKMANN.BENJAMIN@MAYO.EDU
Gregory Worrell WORRELL.GREGORY@MAYO.EDU
David Jones JONES.DAVID@MAYO.EDU
Yogatheesan Varatharajah VARATHA2@ILLINOIS.EDU

Abstract

This paper presents a domain-guided approach for learning representations of scalp-electroencephalograms (EEGs) without relying on expert annotations. Expert labeling of EEGs has proven to be an unscalable process with low inter-reviewer agreement because of the complex and lengthy nature of EEG recordings. Hence, there is a need for machine learning (ML) approaches that can leverage expert domain knowledge without incurring the cost of labor-intensive annotations. Self-supervised learning (SSL) has shown promise in such settings, although existing SSL efforts on EEG data do not fully exploit EEG domain knowledge. Furthermore, it is unclear to what extent SSL models generalize to unseen tasks and datasets. Here we explore whether SSL tasks derived in a domain-guided fashion can learn generalizable EEG representations. Our contributions are three-fold: 1) we propose novel SSL tasks for EEG based on the spatial similarity of brain activity, underlying behavioral states, and age-related differences; 2) we present evidence that an encoder pretrained using the proposed SSL tasks shows strong predictive performance on multiple downstream classifications; and 3) using two large EEG datasets, we show that our encoder generalizes well to multiple EEG datasets during downstream evaluations.

1. Introduction

A scalp-electroencephalogram (EEG) is a biosignal modality that non-invasively measures the electrical activity of a large population of cortical neurons via an array of sensors placed on the subject's scalp (Binnie and Prior, 1994). Scalp EEGs are one of the main diagnostic tests in neurology, where the visual identification of abnormal brain activity indicates the potential for neurological disorders. EEGs also play a crucial role in brain-computer interfaces where the same neural signals help decode or predict brain activity (Casson et al., 2010). Despite its ubiquitous use, the time-consuming review of EEGs is still performed by experts because of the complexity of neurophysiological phenomena. At present, the pace of manual expert labeling cannot match the amount of clinical and research EEG data being acquired. Furthermore, manual EEG review results in low inter-rater agreement due to high variability in expert interpretation (Halford et al., 2017; Williams et al., 1985). As a consequence, under the traditional supervised learning regime, an increasing amount of unlabeled EEG data remain under-utilized and algorithms trained on 'noisy' labels show a degradation in performance, increase in model complexity, and difficulty in identifying relevant features (Frénay and Verleysen, 2013; Frénay et al., 2014). These issues create a strong incentive for researchers to develop methods that can learn from the large amounts of raw unannotated EEG corpora already available.

Encouragingly, a variant of unsupervised machine learning, called self-supervised learning (SSL), has shown great promise in settings where, 1) access to labels is limited; 2) the training process is corrupted by label noise; and 3) out-of-domain generalization is desirable (Hendrycks et al., 2019). In the SSL setting, an encoder is trained in a supervised fashion on un-

labeled data by constructing a 'pretext' learning task that predicts known attributes found in the data itself (i.e., self-supervision). Then, this pretrained encoder is used as a feature extractor for 'downstream' tasks of interest (Liu et al., 2021). Examples of commonly used pretext tasks for time series data include 1) learning to predict a portion of future or masked data, 2) learning to predict the temporal order of a sequence of inputs (Jaiswal et al., 2021), or 3) learning to identify signals that are proximal in time versus distal (Banville et al., 2021). Intuitively, such tasks are designed to recover the temporal structure that exists in a time series. The empirical success of SSL in other domains presents a particularly enticing possibility for EEG - can we make encoders learn desirable physiological or pathological features through bespoke pretext tasks? Arguably, constructing pretext tasks for EEG in collaboration with domain experts is a more scalable approach than the expert manual review of lengthy EEG recordings. In that context, this study investigates two hypotheses:

- SSL can explicitly encode known physiological patterns of scalp EEG data that we desire our model to be sensitive to.
- Physiologically meaningful SSL pretext tasks can enable the learning of EEG representations that transfer (i.e., generalize) seamlessly to multiple downstream tasks and datasets.

To that end, this paper introduces three novel pretext tasks in normal subjects that leverage, 1) spatial similarities across the left and right brain hemispheres, 2) brain's behavioral states, and 3) age-related changes in brain activity (Section 4). We trained an encoder using the above tasks under a multi-task learning setting and evaluated its performance on several downstream tasks and datasets that included subjects with abnormal EEGs. We utilized the TUH EEG corpus (Obeid and Picone, 2016) for encoder training and validation, and the MPI LEMON corpus (Babayan et al., 2019) for out-of-sample evaluation. We fine-tuned the encoder and evaluated its performance on several downstream tasks including, EEG grade (normal, abnormal), eye state (eyes open, eyes closed), age (young, old), and gender (male, female) classifications. Our results (Section 6) indicate that multi-task domain-guided SSL pretraining is effective in learning desirable properties and the learned representations show strong predictive performance and generalization across new subjects, multiple tasks, and out-of-sample datasets.

2. Related Work

Recently, there has been a sharp surge in studies that aim to apply SSL to physiological time series data, ranging from cardiac signals (Kiyasseh et al., 2021; Chen et al., 2021), electronic health record (EHR) timeseries (McDermott et al., 2020; Yèche et al., 2021), to EEG data (Mohsenvand et al., 2020; Banville et al., 2021; Kostas et al., 2021). In the following, we summarize two EEG-based SSL studies closest to our work and highlight the differences.

Mohsenvand et al. (2020) proposed a contrastive learning approach for EEG data adapted from domains such as computer vision that enforced the encoder to be invariant to physiologically plausible data augmentations such as time shift, scaling, masking, filtering, and additive noise. Banville et al. (2021) proposed a 'relative positioning' pretext task which ensured that nearby EEG epochs (epochs denote short segments of EEG) have similar representations and distant epochs have dissimilar representations. A limitation of this pretext task is that distant epochs of brain activity can be very similar if they represent the same behavioral state. Furthermore, both studies trained separate encoder models for each downstream task and dataset. As such, the transfer capabilities of those models to other datasets remain unknown.

Our work differs from the above studies in several ways. First, we propose pretext tasks that leverage EEG's spatial similarities, dynamic behavioral states of the subject, as well as age-related differences in EEG. Second, we propose an approach for estimating a measure of behavioral state from raw data and define more interesting notions of contrast across multiple subjects (as in (Kiyasseh et al., 2021)). In prior work, the definition of contrasting epochs disregards EEG behavioral states and in most cases has been restricted to EEG epochs drawn from the same recording. Third, we evaluate the transfer capability of our approach using an out-of-sample dataset. Specifically, we evaluate whether the representations learned by an encoder model trained on one dataset can generalize to multiple downstream tasks in an out-of-sample dataset with additional fine-tuning.

3. Overall Workflow & Datasets

Figure 1 illustrates the overall workflow of our study. We used the topographical maps in seven frequency bands to represent epoch-level EEG data (described below). We trained an encoder f_θ with a resnet-18

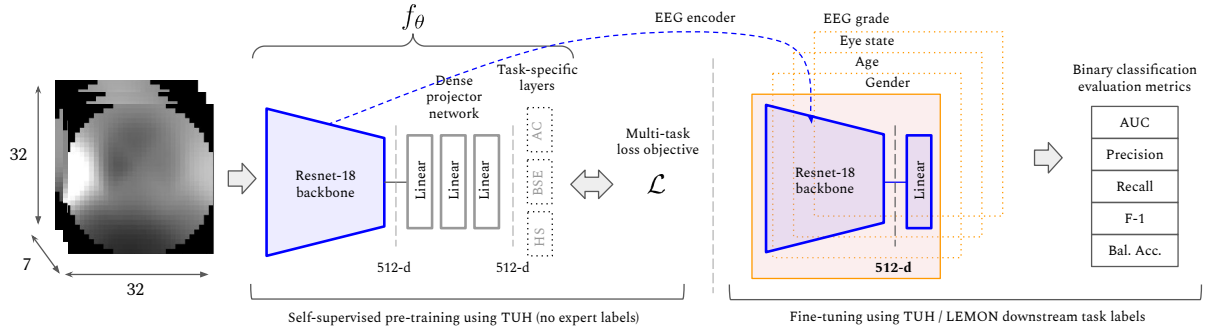


Figure 1: Illustration of the overall workflow. We represent the EEG data as topographical maps, train a resnet-18 encoder (feature extractor) using proposed SSL tasks, and evaluate on multiple downstream tasks after fine-tuning.

backbone (He et al., 2016) using a multi-task loss defined by three SSL tasks (Section 4) using the TUH EEG Corpus. Here the parameters θ represent the parameters of the encoder. We then applied the trained encoder to multiple downstream classification tasks in TUH and LEMON datasets (Section 5). Note that we used only the resnet-18 backbone network (shown in blue in Figure 1) for downstream fine-tuning and that we fine-tuned the same pretrained backbone for each of the four downstream tasks. Section 6 describes the results of our experiments.

Datasets: Our experiments make use of two EEG corpora: 1) the Temple University Hospital (TUH) EEG Corpus (Obeid and Picone, 2016), which contains clinical EEG recordings of patients with neurological disorders, and 2) the Max Planck Institute Leipzig Mind-Brain-Body (LEMON) Dataset (Babayan et al., 2019), which contains resting-state recordings from healthy participants.

TUH EEG: This dataset comprises ∼30,000 EEG recordings collected at TUH starting from 2002. A subset of recordings in TUH EEG have been broadly annotated by experts as either "normal" or "abnormal", and have been released as a derived dataset called the TUH EEG Abnormal Corpus (TUAB). For our experiments, we only utilize the TUAB recordings, leading to a total of 2993 EEGs from 2329 distinct patients. We extracted the age and gender of recorded subjects, assumed here as non-expert labels, from text reports accompanying the EEGs (Rawal and Varatharajah, 2021). Recordings where this extraction failed were discarded.

MPI LEMON: This dataset represents a cross-sectional sample of healthy individuals from Leipzig, Germany. The sample comprised two age groups: young adults (ages 20-35) and older adults (ages 59-77). EEG recordings were acquired using 62 electrodes in the 10-10 sensor configuration with a sampling rate of 2500Hz, for a total of 216 participants. Each subject's session is made up of 16 trials, each 60 seconds long: 8 eyes-closed and 8 eyes-open. We included data from both trials in our experiments. The raw data were corrupted for 4 subjects, which resulted in a useful set of healthy EEGs from a total of 212 subjects. Age group and gender labels are released with the EEGs (78 females and 134 males).

Data preprocessing[1]**:** The EEG preprocessing steps applied to both datasets are as follows: (1) we selected 19 EEG channels, namely 'Fp1', 'Fp2', 'F3', 'F4', 'C3', 'C4', 'P3', 'P4', 'O1', 'O2', 'F7', 'F8', 'T7', 'T8', 'P7', 'P8', 'Fz', 'Cz', and 'Pz', (2) we resampled the recordings to 80Hz, followed by (3) a bandpass filter between 1Hz and 39.5Hz, then (4) we divided the recordings into contiguous non-overlapping epochs of 10-seconds each, and finally (4) we identified and removed bad epochs when the total power in their 'Cz' channel exceeded 2 standard deviations as calculated from statistics of the entire recording.

Data representation: We represented the EEG epochs by 2D images (each of size 32 × 32) that depict the topographical map (or 'topomap') of the spectral power in a brain rhythm band. Specifically, we computed these maps of relative power distribution within 7 frequency bands defined as: delta (2-4Hz), theta (4-8Hz), lower alpha (8-10Hz), higher alpha (10-13Hz), lower beta (13-16Hz), higher beta

1. Based on Makoto's EEG preprocessing pipeline: https://sccn.ucsd.edu/wiki/Makoto's_preprocessing_pipeline

(16-25Hz), and gamma (25-40Hz). Collectively, these topomaps result in a 7-channel image representing an EEG epoch (Figure 1). We min-max scaled these images within each channel such that the values lie between 0 and 1.

4. Model Description

In this section, we describe the proposed SSL tasks (shown in Figure 2), their mathematical formulations and corresponding loss functions, and how they are combined to form a multi-task objective for pretraining. Lastly, we describe the architectural elements of the encoder trained using that objective.

Hemispheric symmetry (HS) task: In general, healthy brain activity is similar across the left and right hemispheres. Therefore, despite naturally arising minor distortions, we would want the feature representations of brain activity derived from both the hemispheres to be similar as well. This forms the basis for our first SSL task. In the following, we use $X \in \mathbb{R}^{7\times32\times32}$ to denote a topomap of brain activity and $X^{aug} \in \mathbb{R}^{7\times32\times32}$ to denote an augmented version of the topomap. We use $Z \in \mathbb{R}^d$ to denote the output of the feature extractor f_θ, where d is the output dimensionality, i.e., $Z = f_\theta(X)$.

Suppose that we obtain X^{aug1} by randomly flipping X horizontally and X^{aug2} by randomly adding Gaussian noise[2]. Our goal is to define a loss that penalizes learning different representations for X^{aug1} and X^{aug2}. To achieve this invariance, we employ the Barlow Twins (BT) loss function (Zbontar et al., 2021). Given a batch of inputs X^{aug1} and X^{aug2}, the construction of the BT loss encourages the model f_θ to learn representations Z^{aug1} and Z^{aug2} such that the cross-correlation matrix between them is close to an identity matrix. The cross-correlation matrix C over a batch of inputs b is given by Eq. 1.

$$\mathcal{C}_{ij} = \frac{\sum_b z^{aug1}_{b,i} z^{aug2}_{b,j}}{\sqrt{\sum_b \left(z^{aug1}_{b,i}\right)^2}\sqrt{\sum_b \left(z^{aug2}_{b,j}\right)^2}} \tag{1}$$

Given C, the hemispheric symmetry loss $\mathcal{L}_{HS}$ is computed using Eq. 2.

$$\mathcal{L}_{HS} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda\left(\sum_i \sum_{j\neq i} \mathcal{C}^2_{ij}\right) \tag{2}$$

2. Although we could simply focus on the similarities between X and X^{aug1}, we found that adding some noise to the original EEG epoch X improves training.

Both terms in the definition serve a distinct purpose. The first term encourages the diagonal elements of C to be closer to 1, thereby making the representations similar for input distortions. The second term forces the off-diagonal terms of C towards 0, thereby decorrelating the embeddings and as a result, preventing the model from learning trivial solutions. The hyperparameter λ balances the relative importance of each term. As discussed in (Tsai et al., 2021), we set λ to $1/d$ so that both terms are weighted equally during optimization.

Behavioral state estimation (BSE) task: The dynamic nature of neural activity is modulated by the behavioral state of the subject. Therefore, it is clearly beneficial for EEG encoders to generate representations sensitive to the behavioral state of the subject. The slow-wave (δ) to fast-wave (β) spectral power ratios in the central brain regions provide robust estimates of the behavioral state of the subject and their attentional control (Kremen et al., 2017). We used the delta-beta power ratio as the proxy measure to construct a pretext task. For an EEG epoch X, the delta-beta ratio (DBR) is defined by Eq. 3.

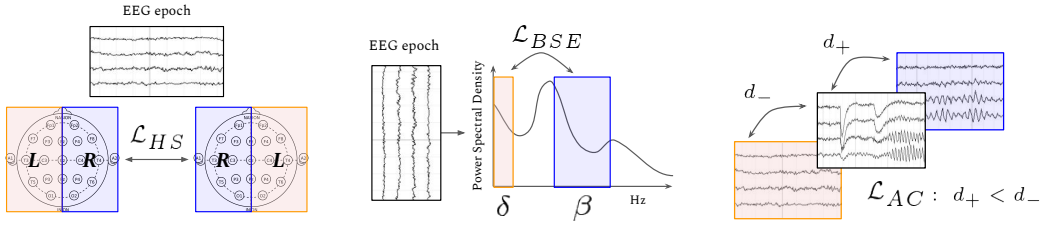
$$\text{DBR}(X) = \frac{p_\delta(X)}{p_\beta(X)} \tag{3}$$

Here, $p_\delta(X)$ and $p_\beta(X)$ are the power of δ and β bands calculated from the original EEG time series corresponding to X. We chose δ band as 2 – 4 Hz and β band as 13 – 25 Hz, and used the C3 and C4 channels to calculate the power ratio. Although DBR is only a proxy measure of the subject's behavioral state, it has the advantage of being easily computable from data. We formulate this task as regression with a mean square loss defined by Eq. 4.

$$\mathcal{L}_{BSE} = \frac{1}{|b|}\sum_b \Big\| g(Z) - \text{DBR}(X) \Big\|^2_2 \tag{4}$$

Here, $g(.)$ takes the representation Z and outputs a prediction of DBR (i.e., a regression head).

Age contrastive (AC) task: Due to the importance of age in modulating brain activity (Rossini et al., 2007), it is desirable for an EEG feature extractor to learn representations that are sensitive to the age-related slowing of EEG. Broadly, the representations of brain activity of younger subjects should look different from those of older subjects. This property can be readily formulated within the contrastive triplet learning paradigm, originally proposed by (Weinberger and Saul, 2009). A triplet is a



(a) Hemispheric symmetry (b) Behavioral state (c) Age contrast

Figure 2: A schematic representation of the three self-supervised learning (SSL) tasks proposed in this study.

training tuple constructed from 3 EEG epochs: 1) an anchor epoch X, 2) an epoch 'similar' to the anchor X_+, and 3) an epoch 'dissimilar' to the anchor X_-. The objective is to learn a feature space in which positive pairs are represented closer than negative pairs, per some measure of similarity s. Given a set of training triplets (X, X_+, X_-), the triplet loss (Schroff et al., 2015), is defined by Eq. 5.

$$\mathcal{L}_{AC} = \max(s(X, X_+) - s(X, X_-) + \gamma, \ 0) \qquad (5)$$

Here, γ is a hyper-parameter called margin.

We define the notion of contrast using two broad age groups: young (age ≤ 40) and old (age ≥ 60), and use the Euclidean distance as a measure of similarity. Two EEG epochs are labeled similar if they come from subjects within the same age group. Conversely, they are labeled dissimilar if they come from different age groups. To make this age contrast physiologically meaningful, we require that all epochs within the triplet are in the same behavioral state (i.e., similar DBR) as the anchor epoch. DBR similarity is determined by the quartiles formed by the distribution of DBR values in the training set.

Model architecture: The overall model architecture and multi-task training scheme are shown in Figure 1. The feature extractor f_θ consists of a resnet-18 'backbone' network followed by three linear layers called the 'projector' network. Due to the presence of this projector network, the last linear layer of the resnet-18 backbone is disabled. This architecture is a scaled-down version of the model used in (Zbontar et al., 2021). At the end, there are task- or loss-specific layers. Specifically, the HS loss needs a batch normalization layer before the loss is computed, while the BSE loss needs an additional linear layer that outputs the DBR prediction for the sample (i.e., g).

The encoder model f_θ accepts a 7-channel topomap and outputs a 512-dimensional embedding. While the full model f_θ is pretrained using the multi-task loss objective $\mathcal{L}$ (Eq. 6), only the resnet-18 backbone is extracted after pretraining and used to fine-tune on the downstream tasks.

Multi-task training: The weights of the network are shared between the three tasks and all three tasks are used to train the encoder simultaneously. Eq. 6 defines the total loss $\mathcal{L}$.

$$\mathcal{L} = \mathcal{L}_{\text{HS}} + \mathcal{L}_{\text{BSE}} + \mathcal{L}_{\text{AC}} \qquad (6)$$

5. Experiments & Evaluation Setup

Experiments: Our experiments consist of qualitative and quantitative evaluations of the proposed multi-task training objective and the learned representations. In addition, we conduct ablation experiments to elucidate the contribution of each of the proposed SSL tasks. The overall semi-supervised workflow is depicted in Figure 1.

First, we trained an encoder model on a subset of the TUH dataset and evaluated whether the physiological characteristics enforced by the SSL tasks are reflected in the representations generated by the encoder. We generated embeddings using the trained encoder for the epochs in the validation set and visualized them with ground-truth overlays in a low-dimensional 2D space using the t-SNE algorithm (Van der Maaten and Hinton, 2008).

Second, we conducted a within-sample validation of the encoder by assessing its transfer performance to different subjects and tasks within the same dataset used for training, i.e., the TUH corpus. We consider three downstream binary classification tasks: 1) whether the EEG was normal or abnormal (referred to as EEG grade), 2) whether the subject was young (age ≤ 45) or old (age > 45; referred to as age), and 3) whether the subject was male or female (referred to as gender). We fine-tuned the encoder on a training set of the TUH dataset and evaluated the classifi-

cation performance on a held-out test set. Third, we conducted an out-of-sample evaluation of the encoder by assessing its transfer performance in the LEMON dataset. We consider three downstream binary classification tasks: 1) whether the subject's eyes were open or closed (referred to as eye state), 2) whether the subject was young (age ≤ 55) or old (age > 55; referred to as age), and 3) whether the subject was male or female (referred to as gender). We fine-tuned the encoder trained on the TUH training dataset on a subset of the LEMON dataset and evaluated the classification performance on a held-out test set in LEMON.

Training and evaluation: We used the stochastic gradient descent optimizer with momentum set to 0.9 for all model training (Qian, 1999). We controlled the learning rate by a policy that oscillates between 0.5 to 0.0001 (Smith, 2017) and dampened the oscillations over time by an exponentially decreasing scaling factor of 0.5. During transfer experiments, we extracted the resnet-18 backbone of the best-fit model pretrained on the TUH data and added a linear output layer for binary classification on downstream tasks. We separately fine-tuned the resnet-18 backbone from the same pretrained model for each of the downstream tasks. We also used a weighted random sampling approach to balance the target classes during fine-tuning.

Data splits: First, we divided the TUH dataset (subjects: 2,328; epochs: 409,083) into train (subjects: 2,095; epochs: 365,483) and validation (subjects: 233; epochs: 43,600) sets for pretraining. For fine-tuning and evaluation on downstream tasks, we performed a 3-way data split including train, validation, and test sets. Table 1 shows those divisions in the TUH and LEMON datasets. We made all data splits using disjoint sets of subjects.

	TUH		LEMON	
	Subjects	Epochs	Subjects	Epochs
Train (∼55%)	1,303	227,728	118	10,487
Valid (∼15%)	326	56,543	30	2,695
Test (∼30%)	699	124,812	64	5,744

Table 1: Train, validation, and test splits.

Deriving subject-level predictions: Because we performed model training using epochs of EEG data, we aggregated epoch-level predictions to form subject-level predictions in cases where the target attribute is at subject-level (grade, age, or gender). We made a simplifying assumption that the signal in each 10s EEG epoch is independent of other epochs in the dataset. We then used a maximum likelihood estimation based on the classifier output $Y_n \in [0, 1]$, which represents the probability that the n^{th} epoch belongs to the positive class. We model the epoch-level predictions of a subject S_i as independent observations made from a Bernoulli trial with an unknown probability π_i, where π_i is the probability that the subject S_i belongs to the positive class. Then, an estimate of π_i that maximizes the likelihood function $\prod_{n=1}^{N(i)} \pi_i^{Y(n)}(1-\pi_i)^{(1-Y(n))}$ after N epochs is given as $\hat{\pi_i} = \frac{\sum_{n=1}^{N} Y_n}{N}$.

Metrics: We used the area under the receiver operating characteristic curve (AUC), precision (Prec.), recall (Rec.), F1 score, and balanced accuracy (B.Acc.) to evaluate downstream classifications. We chose an optimal decision threshold using Youden's J statistic (Youden (1950)) to calculate precision, recall, F1, and balanced accuracy scores for the held-out test sets. We evaluated the final models on the held-out test set in a leave-one-subject-out fashion to obtain multiple AUC scores for each model. Table 2 displays the mean AUC scores and standard deviations obtained in those evaluations (see the appendix for a full report of all the metrics).

Baselines comparisons: We compared the transfer performance of our proposed method to: 1) a purely supervised linear classification baseline, and 2) the temporal contrastive SSL pretext task called 'relative positioning' (RP) (Banville et al., 2021). We trained the linear classifier on a flattened feature input of power spectral density values of the same 19 channels used to generate the topographic maps. We used a grid search to choose the values of the regularization strength and elastic-net mixing ratio. For the relative positioning benchmark, we computed the channel-wise normalized time series and ShallowNet model as described in the original paper. We heuristically set the hyperparameters τ_{pos} and τ_{neg} to 6 and 12 respectively, meaning that the positive context for the anchor spans 60s (30s on each side of the anchor), while the negative context lies beyond 60s on either side of the anchor. From each recording, we sampled 1000 positive and 1000 negative training tuples during pretraining. We performed downstream evaluations in a similar fashion as our proposed model described above. Finally, we used the two-sample Kolmogorov-Smirnov test to determine whether the AUC scores obtained from leave-one-out evaluation of the best baseline and those from the best ablated model are drawn from the same distribution.

Software & hardware: The preprocessing, feature extraction, and experiments were implemented using a combination of the following Python libraries: 1) scikit-learn (Pedregosa et al., 2011), 2) MNE-python (Gramfort et al., 2013), 3) PyTorch (Paszke et al., 2019), 4) NumPy (Harris et al., 2020), and 5) Braindecode (Schirrmeister et al., 2017). The experiments were performed using 2 Nvidia RTX 3090 GPUs.

6. Results

We summarize our results in Figure 3 and Table 2. Figure 3 illustrates the qualitative experiments and Table 2 provides a comparison between our encoders and baseline models in both TUH and LEMON datasets (based on AUC metric alone). The appendix includes additional metrics such as precision, recall, F-1, and balanced accuracy.

Visualization of the feature space: Figure 3 illustrates how the SSL tasks affect the embeddings generated by the encoder. Figures 3(a) and 3(b) show the generated embeddings when the encoder was trained with all three SSL tasks, and BSE and AC tasks alone, respectively. Here, the green points indicate representations of unaugmented data and the red points indicate those of hemispherically-flipped data in the same low-dimensional 2D space denoted by $z1$ and $z2$. We observe that adding the HS SSL task enables the learning of representations that are unaffected by the spatial flipping of the input. Figure 3(c) shows the predicted DBR values (using the linear layer $g(.)$) and ground truth DBR values when the encoder was trained with BSE task alone. We observe that the predicted DBR values and ground-truth computed DBR values align well. Lastly, Figure 3(d) shows the embeddings generated from epochs of young (red) and old subjects (green) when the encoder was trained using the AC task alone. Here, we fixed the epochs to be in a specific behavioral state because behavioral states can confound age-related differences. We observe a stronger separation between subjects in contrasting age groups than the subjects within the same age group. Overall, our results suggest that the EEG representations generated using the proposed SSL framework reflect desirable physiological properties.

Within-sample transfer evaluation: The first set of columns in Table 2 show the results obtained after fine-tuning the pre-trained encoder on TUH task labels. We observe that the full multi-task setting does not always provide the best performance. Nonetheless, ablated versions substantially outperform the linear baseline in all the tasks - EEG grade (0.88 vs. 0.92 AUC), age (0.74 vs. 0.80 AUC), and gender (0.68 vs. 0.70 AUC). Interestingly, the ablated version trained with BSE SSL task alone provides the best performance in each case.

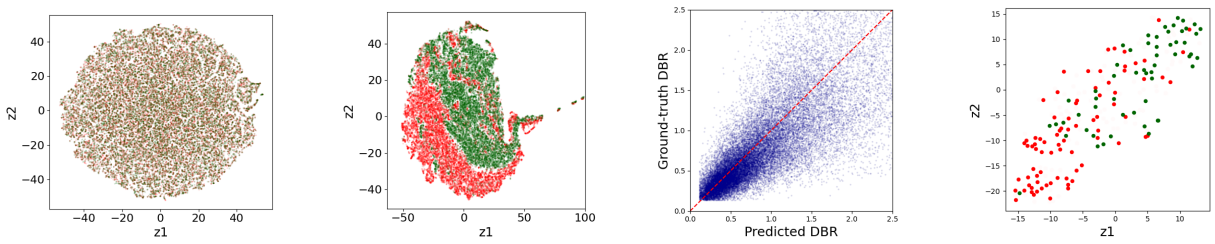
Out-of-sample transfer evaluation: We utilized the LEMON corpus to assess the pre-trained models' transfer capabilities to domains or environments not seen during training. The evaluation procedure remains identical to that of the internal validation. The second set of columns in Table 2 show the results obtained after fine-tuning the pre-trained encoder on LEMON task labels. We observe that our method performs competitively to the linear baseline in the eye state task (0.89 vs. 0.89 AUC) but substantially outperforms in the age (0.95 vs. 0.99 AUC) and gender (0.65 vs. 0.80 AUC) tasks.

Comparison with state-of-the-art: Results obtained for an independent implementation of the temporal-domain 'relative positioning' (RP) pretext task are listed as the ShallowNet model in Table 2. RP performs better at age and gender prediction on TUH, and on eye state prediction on LEMON. Our method performs better on age and gender prediction on LEMON while performing competitively to RP in EEG grade prediction on TUH. Broadly, our multi-task training based on physiologically inspired SSL tasks performs better than this benchmark on transfer tasks evaluated on out-of-sample data.

Ablation studies: Overall, models trained using all three tasks together show performance that is at the least competitive with the linear baseline and in most cases, significantly exceeds it. However, the performance of ablated versions is more nuanced. Interestingly, it is not the case that more self-supervision tasks (i.e., two or more tasks) perform better than a single task. On the other hand, while BSE task alone provides the best within-sample performance (TUH), it does not perform equally well on out-of-sample data (LEMON).

7. Discussion

In this paper, we studied whether EEG encoders trained on a large unlabeled dataset using domain-guided self-supervision can help improve downstream classification performance in small labeled datasets. We utilized a resnet-18 encoder with EEG data represented as topographical maps and derived three SSL tasks inspired from 1) hemispheric similarities in



(*a*) HS + BSE + AC (*b*) BSE + AC (*c*) BSE (*d*) AC

Figure 3: Visualizations showing how the SSL tasks affect the embeddings generated by the encoder. The subtitles denote the SSL tasks used in encoder pretraining. Note that we fixed the epochs in 3(*d*) to be in a specific behavioral state because behavioral states can confound age-related differences.

Models	TUH			LEMON		
	EEG Grade	Age	Gender	Eye State	Age	Gender
Linear	0.884 (4e-4)	0.741 (6e-4)	0.675 (6e-4)	0.888 (2e-3)	0.949 (3e-3)	0.646 (1e-2)
ShallowNet	0.909 (3e-4)	**0.872 (4e-4)***	**0.783 (1e-3)***	**0.948 (1e-3)***	0.945 (3e-3)	0.721 (9e-3)
HS only	0.910 (3e-4)	0.771 (1e-3)	0.648 (1e-3)	0.885 (2e-3)	0.970 (2e-3)	0.740 (8e-3)
BSE only	**0.918 (3e-4)***	0.797 (1e-3)	0.696 (1e-3)	0.892 (2e-3)	0.972 (2e-3)	0.725 (9e-3)
AC only	0.912 (3e-4)	0.792 (1e-3)	0.652 (1e-3)	0.870 (2e-3)	0.970 (2e-3)	0.703 (1e-2)
HS-BSE	0.914 (3e-4)	0.787 (1e-3)	0.668 (1e-3)	0.885 (2e-3)	0.983 (1e-3)	0.693 (9e-3)
HS-AC	0.895 (4e-5)	0.723 (1e-3)	0.636 (1e-3)	0.861 (2e-3)	0.907 (5e-3)	0.649 (9e-3)
BSE-AC	0.913 (3e-4)	0.784 (1e-3)	0.675 (1e-3)	0.878 (2e-3)	**0.987 (1e-3)***	0.704 (9e-3)
HS-BSE-AC	0.907 (3e-4)	0.748 (1e-3)	0.638 (1e-3)	0.870 (2e-3)	0.984 (2e-3)	**0.803 (8e-3)***

Table 2: Within-sample (TUH) and out-of-sample (LEMON) evaluation of ablated versions of the proposed approach (last seven rows) and baselines (first two rows). Results shown are AUC values obtained on the held-out set of TUH and LEMON subjects on multiple classification tasks, TUH: 1) 'EEG grade': normal vs. abnormal, 2) 'Age': young vs. old, and 3) 'Gender': male vs. female; LEMON: 1) 'Eye State': eyes open vs. eyes closed, 2) 'Age': young vs. old, and 3) 'Gender': male vs. female. * indicates a statistically significant result ($p < 0.01$).

brain activity, 2) behavioral states, and 3) age-related EEG changes. We fine-tuned this encoder and evaluated its performance on several downstream tasks in both within-sample and out-of-sample EEG recordings. Our results indicate that the proposed encoder performs substantially better than fully-supervised linear models on both within-sample and out-of-sample experiments and competitively with the current state-of-the-art SSL model with slightly superior performance in out-of-sample tasks.

SSL pretext tasks: Our major contribution is the design of pretext tasks inspired by neurophysiological domain knowledge such as hemispheric similarity of brain activity, behavioral states of the subject, and age-related changes. Our hypothesis was that training a feature encoder using these properties can enable the learning of generalizable features that indicate irregular/pathological asymmetries, behavioral state anomalies, and changes orthogonal to normal aging, which can then help with downstream classifications. Our results also indicate that the supervision provided by the proposed SSL tasks does help with learning generalizable representations. While our study establishes the feasibility of deriving domain-specific SSL tasks for EEG data, there is scope to further increase the granularity and/or quality of self-supervision. For example, subjects between the ages of 40 and 60 could be included to learn finer age-related changes. Additionally, mixing frequency bands other than delta and beta can better estimate behavioral state than our current choice. Future work can also design self-supervision to integrate idiosyncratic asymmetries and pathological changes caused by neurological disorders such as epilepsy and Alzheimer's. Readily computable measures of brain connectivity, as well as information from EEG text reports, remain untapped sources of expert domain guidance that do not need epoch-level annotations.

Generalization: EEG is a highly non-stationary signal i.e., its statistics evolve over time. As a consequence, there is significant inter-subject and inter-session variability in the signal recorded and generalization remains a challenge for predictive EEG models (Roy et al., 2019). To evaluate whether domain-guided self-supervision can partly address this challenge, we pretrained an encoder on large amounts of clinically acquired data (TUH EEG) and validated its performance on a smaller research-grade dataset (LEMON). Our results indicate that domain-guided SSL does provide competitive performance (in some cases, superior) on unseen tasks and out-of-sample datasets. The input representation of topomaps (Section 3) limits our method's ability to transfer to low-density EEG datasets, such as those found in sleep staging and brain-computer interfacing. Since topomaps are computed using spatial interpolation, a spatially low density of EEG sensors will result in an overly smoothed input. Building general EEG feature extractors that are robust to low-density EEG acquisitions remains an area for future work.

Multi-task training: As can be seen from Table 2, it is difficult to suggest which pretext task will prove beneficial for a particular downstream task of interest. However, our results indicate that combining multiple types of pretext tasks is very likely to improve performance over a linear baseline. The fact that the temporal RP task outperforms in certain cases is additional evidence in support of multi-task training. Therefore, we argue that multi-task SSL pretraining is an essential component in the development of general-purpose EEG encoders. However, we also observed that in certain multi-task combinations (HS-AC, for example), one task destabilizes the training of another, leading to worse downstream performance. In such cases, a biased weighting scheme, as opposed to the equal weighting done in our experiments, is likely to be a promising approach.

Backbone model: Our choice of a resnet-18 backbone was motivated by the image-like representation of topomaps and by existing studies that found convolutional and residual networks to be suitable for the classification of physiological signals (Hannun et al., 2019; Faust et al., 2018; Schirrmeister et al., 2017). However, it is plausible that a more physiologically-driven architecture such as ShallowNet (Schirrmeister et al., 2017) trained with the proposed SSL tasks could provide superior results.

Broader impact: Conventional analysis of EEG relies on manual expert annotations which are limited due to their time-consuming nature, inherent signal complexity, and low inter-rater agreement. Through the design of bespoke pretext tasks, we can potentially alleviate the label-centric nature of large-scale EEG predictive modeling (Roy et al., 2019). By avoiding expert annotations, well-designed SSL pretext tasks could also unlock novel clinical applications in scenarios where deep domain knowledge already exists but labeled data does not.

Code & data availability: The datasets used in this study are already publicly available. The model definitions, code to train the proposed pretext tasks, and evaluation scripts needed to reproduce results in Table 2 and Supplementary Tables 3 and 4 are available at https://github.com/neerajwagh/eeg-self-supervision.

8. Conclusion

This study introduced a representation learning method that leverages self-supervised learning (SSL) to learn physiologically relevant features from scalp EEG data without expert annotations. We proposed three pretext tasks derived from healthy EEG patterns that aim to exploit: 1) the hemispheric similarities of brain activity, 2) behavioral states, and 3) age-related EEG changes. Experiments indicate that an encoder pretrained using the proposed method exhibits competitive predictive performance in downstream classification tasks and strong generalization across subjects, tasks, and datasets. Our future efforts will investigate additional domain-guided SSL tasks, different backbone models, and more clinically-oriented downstream tasks. Because SSL enables the integration of domain expertise without incurring the cost of labor-intensive annotations, it has the potential to unlock novel clinical and consumer EEG applications, for which expert knowledge exists but annotated data does not.

9. Acknowledgements

This research was supported by the National Science Foundation (Award No. IIS-2105233) and the Mayo Clinic Neurology Artificial Intelligence Program.

References

Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D Reinelt, Andrea MF Reiter, Josefin Röbbig, H Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, et al. A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults. *Scientific data*, 6(1):1–21, 2019.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

C D Binnie and P F Prior. Electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry*, 57(11):1308–1319, 1994. ISSN 0022-3050. doi: 10.1136/jnnp.57.11.1308. URL https://jnnp.bmj.com/content/57/11/1308.

Alexander J. Casson, David C. Yates, Shelagh J.M. Smith, John S. Duncan, and Esther Rodriguez-Villegas. Wearable electroencephalography. *IEEE Engineering in Medicine and Biology Magazine*, 29 (3):44–56, 2010. doi: 10.1109/MEMB.2010.936545.

Brian Chen, Golara Javadi, Amoon Jamzad, Alexander Hamilton, Stephanie Sibley, Purang Abolmaesumi, David Maslove, and Parvin Mousavi. Detecting atrial fibrillation in icu telemetry data with weak labels. 2021.

Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine*, 161:1–13, 2018.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Benoît Frénay, Ata Kabán, et al. A comprehensive introduction to label noise. In *ESANN*. Citeseer, 2014.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7: 267, 2013.

Jonathan J Halford, Amir Arain, Giridhar P Kalamangalam, Suzette M LaRoche, Bonilha Leonardo, Maysaa Basha, Nabil J Azar, Ekrem Kutluay, Gabriel U Martz, Wolf J Bethany, et al. Characteristics of eeg interpreters associated with higher interrater agreement. *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, 34(2):168, 2017.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25 (1):65–69, 2019.

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person reidentification. *arXiv preprint arXiv:1703.07737*, 2017.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2021.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *arXiv preprint arXiv:2101.12037*, 2021.

Vaclav Kremen, Juliano J Duque, Benjamin H Brinkmann, Brent M Berry, Michal T Kucewicz, Fatemeh Khadjevand, Jamie Van Gompel, Matt Stead, Erik K St Louis, and Gregory A Worrell. Behavioral state classification in epileptic brain using intracranial electrophysiology. *Journal of Neural Engineering*, 14(2):026001, jan 2017. doi: 10.1088/1741-2552/aa5688. URL https://doi.org/10.1088/1741-2552/aa5688.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021. doi: 10.1109/TKDE.2021.3090866.

Matthew McDermott, Bret Nestor, Evan Kim, Wancong Zhang, Anna Goldenberg, Peter Szolovits, and Marzyeh Ghassemi. A comprehensive evaluation of multi-task learning and multi-task pretraining on ehr time-series data. *arXiv preprint arXiv:2007.10185*, 2020.

Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health*, pages 238–253. PMLR, 2020.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1): 145–151, 1999.

Sam Rawal and Yogatheesan Varatharajah. Score-it: A machine learning-based tool for automatic standardization of eeg reports. *Machine Learning for Healthcare (Clinical Abstract)*, 2021.

Paolo M. Rossini, Simone Rossi, Claudio Babiloni, and John Polich. Clinical neurophysiology of aging brain: From normal aging to neurodegeneration. *Progress in Neurobiology*, 83(6):375–400, 2007. ISSN 0301-0082. doi: https://doi.org/10.1016/j.pneurobio.2007.07.010. URL https://www.sciencedirect.com/science/article/pii/S0301008207001451.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

Yao-Hung Hubert Tsai, Shaojie Bai, Louis-Philippe Morency, and Ruslan Salakhutdinov. A note on connecting barlow twins with negative-sample-free contrastive learning. *arXiv preprint arXiv:2104.13712*, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

George W. Williams, Hans O. Lüders, Abraham Brickner, Marlene Goormastic, and Donald W. Klass. Interobserver variability in eeg interpretation. *Neurology*, 35(12):1714–1714, 1985. ISSN 0028-3878. doi: 10.1212/WNL.35.12.1714. URL https://n.neurology.org/content/35/12/1714.

Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. Neighborhood contrastive learning applied to online patient monitoring. *arXiv preprint arXiv:2106.05142*, 2021.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

Appendix A. Supplementary Results

Below are extended versions of Table 2 with additional metrics.

Models	EEG Grade					Age					Gender				
	AUC	Prec.	Rec.	F-1	B.Acc.	AUC	Prec.	Rec.	F-1	B.Acc.	AUC	Prec.	Rec.	F-1	B.Acc.
Linear	0.88	0.78	0.89	0.83	0.81	0.74	0.61	0.70	0.65	0.69	0.67	0.58	0.75	0.65	0.63
ShallowNet	0.91	0.82	0.89	0.85	0.84	0.87	0.77	0.79	0.78	0.81	0.78	0.73	0.7	0.72	0.74
HS only	0.91	0.83	0.86	0.85	0.84	0.77	0.64	0.72	0.68	0.72	0.65	0.58	0.69	0.63	0.63
BSE only	0.92	0.83	0.90	0.86	0.85	0.80	0.66	0.75	0.70	0.74	0.70	0.63	0.68	0.65	0.66
AC only	0.91	0.83	0.87	0.85	0.84	0.79	0.66	0.75	0.70	0.74	0.65	0.59	0.68	0.63	0.63
HS-BSE	0.91	0.83	0.90	0.86	0.85	0.79	0.65	0.76	0.70	0.73	0.67	0.58	0.75	0.66	0.64
HS-AC	0.89	0.81	0.87	0.84	0.83	0.72	0.61	0.64	0.62	0.68	0.64	0.57	0.73	0.64	0.62
BSE-AC	0.91	0.82	0.90	0.86	0.85	0.78	0.64	0.76	0.70	0.73	0.67	0.58	0.77	0.66	0.64
HS-BSE-AC	0.91	0.84	0.87	0.85	0.85	0.75	0.59	0.77	0.67	0.70	0.64	0.58	0.65	0.61	0.61

Table 3: Within-sample evaluation of ablated versions of the proposed approach and baselines. Results shown are obtained on the complete held-out set of TUH subjects (i.e. without the leave-one-out evaluation scheme followed for Table 2) on three binary classification tasks: 1) 'EEG grade': normal vs. abnormal, 2) 'Age': young vs. old, and 3) 'Gender': male vs. female.

Models	Eye state					Age					Gender				
	AUC	Prec.	Rec.	F-1	B.Acc.	AUC	Prec.	Rec.	F-1	B.Acc.	AUC	Prec.	Rec.	F-1	B.Acc.
Linear	0.89	0.81	0.80	0.81	0.81	0.95	0.97	0.84	0.90	0.89	0.65	0.76	0.65	0.70	0.66
ShallowNet	0.95	0.89	0.86	0.88	0.88	0.94	0.95	0.86	0.9	0.88	0.72	0.82	0.7	0.76	0.72
HS only	0.89	0.83	0.78	0.80	0.81	0.97	1.00	0.84	0.91	0.92	0.74	0.92	0.58	0.71	0.75
BSE only	0.89	0.82	0.79	0.80	0.81	0.97	0.97	0.86	0.91	0.91	0.73	0.77	0.83	0.80	0.70
AC only	0.87	0.76	0.84	0.80	0.79	0.97	0.97	0.88	0.93	0.92	0.70	0.76	0.93	0.83	0.71
HS-BSE	0.88	0.84	0.76	0.80	0.81	0.98	0.98	0.93	0.95	0.94	0.69	0.80	0.58	0.67	0.66
HS-AC	0.86	0.75	0.84	0.79	0.78	0.91	0.95	0.86	0.90	0.88	0.65	0.81	0.63	0.70	0.69
BSE-AC	0.88	0.81	0.78	0.79	0.80	0.99	0.98	0.95	0.96	0.95	0.70	0.81	0.73	0.76	0.72
HS-BSE-AC	0.87	0.81	0.78	0.79	0.80	0.98	1.00	0.91	0.95	0.95	0.80	0.87	0.83	0.85	0.81

Table 4: Out-of-sample evaluation of ablated versions of the proposed approach and baselines. Results shown are obtained on the complete held-out set of LEMON subjects (i.e. without the leave-one-out evaluation scheme followed for Table 2) on three binary classification tasks: 1) 'Eye state': eyes open vs. eyes closed, 2) 'Age': young vs. old, and 3) 'Gender': male vs. female.

Appendix B. Additional Discussion

Early stopping: Our pretraining procedure utilized an early stopping criterion based on validation loss. However, it is unclear whether the best-fit pretrained model leads to the best downstream performance, as noted in (Banville et al., 2021). In other words, it is possible that an overfitted pretrained model may yield better downstream performance upon fine-tuning. Further experiments are needed to elucidate how overfitting affects model generalizability.

Online triplet mining: The empirical success of triplet-based training depends heavily on the 'hardness' of triplets sampled (Hermans et al., 2017). Given an anchor X and a margin γ, 'hard' triplets are those that the model will find the hardest to learn i.e., tuples where X_+ is maximally away from X and X_- is minimally close or within a distance of γ to X. With each model update, training samples become increasingly 'easier'. Online mining of triplets ensures that the learning task (Eq. 5) continues to remain moderately difficult for the model.