SELF-SUPERVISED LEARNING FOR ENCODING BETWEEN-SUBJECT INFORMATION IN CLINICAL EEG

Sam Gijsen & Kerstin Ritter
Department of Psychiatry and Psychotherapy, Charité - Universitätsmedizin Berlin, Berlin, Germany
Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany
sam.gijsen@charite.de

ABSTRACT

Progress in deep learning for the analysis of clinical EEG data has been hindered by label noise and labeled data sample sizes. While self-supervised learning (SSL) offers a promising solution by learning representations without labels, their practical utility for clinical applications remains poorly understood. Through systematic evaluation using two large clinical EEG datasets, we provide a comprehensive assessment of SSL for pathology detection while controlling for demographic confounds. We introduce a novel yet simple contrastive learning approach that explicitly encodes between-subject information, achieving superior detection of both neurological and psychiatric pathology compared to existing methods. We use data subsampling to highlight differences in the dynamics of representation learning between the neurological and psychiatric domain. Our evaluations help characterize the strengths and limitations of current SSL approaches, thereby providing guidance for applying and developing SSL methods for clinical EEG. Code and pretrained models are available at https://github.com/SamGijsen/SubCLR/.

1 INTRODUCTION

There is a longstanding and rapidly growing need and interest in the clinical use of neural signals for the diagnosis of pathology. Such applications are widely used in critical areas, such as magnetic resonance imaging for degenerative disease (Filippi et al., 2016; Jack Jr et al., 2016; NICE, 2018) and electroencephalography (EEG) for epilepsy (Binnie & Stefan, 1999; Jing et al., 2020) and sleep disorders (Malhotra & Avidan, 2013). While traditional analysis methods rely mostly on expert interpretation, the recent success of deep learning in computer vision and natural language processing has sparked interest in automated diagnostic applications. EEG presents a compelling opportunity due to its mobile, low-cost, and non-invasive nature. However, despite initial optimism, statistical modeling of neuroimaging data has proven challenging, especially in psychiatric domains, limiting its clinical relevance (Marek & Laumann, 2024).

The challenges in clinical neuroimaging applications stem from disease heterogeneity in both neurological disorders (e.g., seizure types, dementia subtypes) and psychiatric conditions (e.g., anxiety, depression variants). This heterogeneity manifests in varying symptoms, severity levels, and neural correlates across subgroups (Moretti et al., 2010; Price et al., 2017; Loo et al., 2018; Newson & Thiagarajan, 2019). Predicting neurological disorders is generally easier than psychological conditions, as neurological disorders like epilepsy can be diagnosed through visual EEG inspection, reflecting distinct signal-to-noise regimes. In contrast, psychological conditions exhibit marked label noise due to low inter-rater reliability and evolving diagnostic criteria (Freedman et al., 2013; Reed et al., 2018), though neurological conditions also face similar challenges (Dubois et al., 2021; Koch et al., 2021). Consequently, framing disease detection as an end-to-end supervised problem may be suboptimal, as models struggle to learn normative data distributions and identify heterogeneous pathological anomalies, further compounded by data noise and limited neuroimaging sample sizes (Gazzar et al., 2022).

Self-supervised learning (SSL) has been posited as a promising method to alleviate multiple of these issues as it allows for the pretraining of models without the use of labels. Particular success

in computer vision has been achieved by a family of methods which train a deep encoder model to be invariant to a set of hand-picked data augmentations which preserve semantic information (Chen et al., 2020; Grill et al., 2020). The performance on downstream tasks has been attributed to the richness of the learned high-dimensional features. Such methods may thus hold promise for modeling the heterogeneity and disease subtypes found in clinical neuroimaging. Furthermore, label noise is circumvented during pretraining by omitting labels, which has the additional benefit of enabling the use of larger sample sizes as unlabeled data becomes available.

Initial studies applying SSL to EEG data have shown promising results, with explorations beyond augmentation-based methods (Mohsenvand et al., 2020; Yang et al., 2021) including those based on the temporal ordering of EEG (Banville et al., 2021) or signal reconstruction (Jiang et al., 2024). Nevertheless, current work on SSL approaches for EEG analysis suffers from several limitations. Studies typically evaluate methods using different pretraining datasets, model architectures, and parameter counts, making it difficult to isolate the impact of the representation learning strategy itself. Moreover, the field lacks systematic comparisons with simpler baselines, leaving open questions about when complex SSL approaches truly add value. Overall, it is challenging to draw conclusions about how to learn meaningful representations of neural data.

This paper presents a systematic analysis of representation learning using SSL for pathology detection in neuroimaging. We compare various SSL approaches against baseline methods while controlling for architecture and training data. To this end, we use the TUAB dataset (Obeid & Picone, 2016) which features mainly neurological pathology and the Healthy Brain Network dataset (Alexander et al., 2017), which contains a variety of psychiatric disorders. Through careful dataset subsampling and controlling for demographic confounds, we investigate four key questions: (1) How do learned representations compare for pathology detection? (2) Can SSL methods differentiate between healthy and pathological cases using unlabeled data? (3) What are the scaling dynamics of SSL methods with dataset size? (4) Can learned representations be transferred to small, external data? (5) Do these representation learning dynamics differ between neurological and psychological domains? As we perform our experiments on two clinical datasets, we provide insights into the practical utility of representation learning for different types of disorders.

2 Related Work

EEG-Based Pathology Detection. The use of machine learning with EEG data for neurological pathology detection has been shown to be effective with accuracies well above 80% (Gemein et al., 2020; Khan et al., 2022). Self-supervised learning was found to be particularly useful when limited labeled data is available (Banville et al., 2021; Gijsen & Ritter, 2024). However, with more labeled data, expert-based features have enabled similar performance as supervised deep learning (Gemein et al., 2020; Kiessner et al., 2024). Given that this is also observed for tasks such as motor imagery decoding (Schirrmeister et al., 2017), general factors such as signal-to-noise ratios may be the cause. However, such domains have very limited data. Meanwhile, for neurological pathology it is a common observation despite significantly larger datasets and a variety of predominantly CNN-based architectures (Roy et al., 2019; Gemein et al., 2020; Western et al., 2021; Kiessner et al., 2024; Darvishi-Bayazi et al., 2024). Multiple authors have posited that label noise in clinical settings may be constraining further improvement (Engemann et al., 2018; Gemein et al., 2020).

The extent to which psychiatric conditions can be predicted based on neuroimaging data is difficult to gauge from the literature, as many of the numerous studies rely on small sample sizes and often include poor evaluation methodology. As a result, prediction accuracies are reported across wide ranges, with negative correlations between model accuracy and study sample sizes indicating inflated reported results (Arbabshirani et al., 2017; Kambeitz et al., 2018; Flint et al., 2021). Whereas this has received more attention in magnetic resonance imaging, this has also been found in EEG (Watts et al., 2022). Although the Healthy Brain Network dataset has made such analyses possible on a larger dataset (Alexander et al., 2017) which even spurred a benchmark competition (Langer et al., 2022), we are not aware of published results.

Self-supervised learning with EEG. Initial studies applying SSL to EEG data have shown promising results. Banville et al. (2021) investigated SSL for pathology detection by developing pretraining tasks relying on the temporal ordering of epochs of EEG data. SSL was shown to outperform supervised baseline models when a considerable subset of labels was withheld. In subsequent work,

augmentation-based SSL was found to be more performative on various tasks, including pathology detection (Mohsenvand et al., 2020). Further literature has studied this type of SSL for EEG demonstrating enhanced label-efficiency, but predominantly for cases of sleep staging (Yang et al., 2021; Rommel et al., 2022), emotion recognition (Zhang et al., 2022b), and motor imagery (Cheng et al., 2020; Rommel et al., 2022). Yet, detailed baseline comparisons are often omitted, complicating attempts to infer the practical utility of such methods for EEG data. These findings highlight the potential utility of SSL in clinical settings, which often face significant shortages of labeled data. Recent work has focused on scaling, leveraging multiple datasets for pretraining in combination with large transformers, often pretrained for signal reconstruction (Jiang et al., 2024; Yang et al., 2024; Dimofte et al., 2025). Finally, a novel avenue is being explored by integrating natural language during pretraining (Gijsen & Ritter, 2024).

3 METHODS

3.1 DATA AND PREDICTION TASKS

We based our analyses on two large EEG datasets with available pathology information. First, the Temple University Hospital Abnormal EEG Corpus (TUAB; (Obeid & Picone, 2016)), which contains clinical EEG data predominantly of adults recorded in a hospital setting. Pathology is described to largely be neurological, including epilepsy, stroke, and Alzheimer's disease, among others (Gemein et al., 2020), although a precise characterization is not available. Each recording was labeled by physicians as normal or pathologically abnormal, which we used as a binary classification prediction target. The dataset comes split into a training (n=2711) and evaluation (n=276) set, of which we use the latter as a hold-out test set.

Second, we used EEG recordings of the Healthy Brain Network (HBN; (Alexander et al., 2017)), which is a clinical, pediatric dataset mainly covering psychiatric disorders. The dataset contains considerable comorbidity, with the most common disorder categories as per the DSM-V being attention-deficit/hyperactivity disorders, anxiety disorders, specific learning disorders, autism spectrum disorders, disruptive disorders, communication disorders, and depression disorders (Langer et al., 2022). For 2707 subjects we downloaded complete resting state EEG data with the required phenotypic and meta data. We sampled 15% of subjects to constitute a hold-out test set, balanced based on age, sex, and diagnoses. This yields a training set of 2300 subjects and a test set of 407 subjects. To enable classification, we construct a binary target which aims to capture the overall level of behavioural functioning. To this end, we relied on the Children's General Assessment Scale (Shaffer et al., 1983), which is commonly used by mental health clinicians and ranges from 1 to 100. While higher scores correspond to better functioning, lower scores often result from one or multiple psychiatric conditions. We perform 'extreme-group' prediction, differentiating between individuals with moderate-to-severe functional impairment versus those with minimal impairment. To simplify notation, we henceforth refer to these groups as pathological and normal respectively, but we stress that in both datasets no subsets can be confidently stated to represent the healthy population due to the clinical setting in which data collection took place.

3.1.1 DATASET SUBSAMPLING AND PREPROCESSING

Since participant sex and age strongly affect EEG signals, models may considerably rely on these (confounding) factors. To ensure models need to rely on information directly related to pathology, we subsample the datasets to match sex and age distributions between pathological and normal groups (Table 1, with precise information in Appendix A.3.1). Whereas this is sometimes performed for downstream evaluation, we importantly also do this for the pretraining data. In Appendix A.3.1), we also detail the EEG preprocessing, which follows literature standards.

3.2 SELF-SUPERVISED LEARNING

The goal of self-supervised learning is to generate data representations useful for downstream tasks. A common approach involves training a deep encoder, such as a convolutional neural network, to become invariant to data augmentations. This method, widely adopted in computer vision, has demonstrated strong performance across various tasks (Chen et al., 2020; Grill et al., 2020). Differences in EEG recording techniques, such as channel counts and montages, make translating between

Table 1: Dataset statistics. The 'SSL samples' column indicates the amount of samples used for SSL pretraining, which operates on single-channels of EEG-crops.

Data subset	EEG files	Crops (10s)	SSL Samples	% Male	Mean Age (std)
TUAB train (PAT+NOR)	928	60K	1.2M	50.0	49.3 (16.8)
TUAB train (NOR)	928	60K	1.2M	50.0	48.8 (16.5)
TUAB test	276	18K	377K	53.6	50.7 (18.3)
HBN train (PAT+NOR)	476	10K	1.1M	66.8	10.9 (3.6)
HBN train (NOR)	476	10K	1.1M	59.2	10.6 (3.3)
HBN test	138	3K	312K	65.2	11.0 (3.3)

datasets challenging, severely limiting flexibility in downstream tasks. To overcome this issue we pretrain a single-channel encoder model, which is agnostic to the channel-layout, as employed in Mohsenvand et al. (2020). For the downstream task, the linear probe then is trained on the concatenated representations of each of the channels. This enables us to evaluate whether learned representations generalize to small external datasets, which is commonly required in clinical contexts. Consequently, the input data for the pretraining task is a batch of single-channel EEG epochs $\mathbf{x}$ and we aim to learn a representation $\mathbf{h}$ to be used for the downstream tasks of binary pathology detection on the TUAB and HBN datasets. For every analysis, we perform five pretraining runs with different random initializations.

We compare various SSL methods by the predictive performance of their representations on downstream tasks. Specifically, we compare to SimCLR (Chen et al., 2020), Bootstrap-Your-Own-Latent (BYOL; Grill et al., 2020), VICReg (Bardes et al., 2021), and Contrast with the World Representation (ContraWR; Yang et al., 2022a). For a description of these methods, please see Appendix A.3.2. We describe the used data augmentations in Appendix A.3.3.

3.2.1 CONTRASTIVE LEARNING WITHOUT AUGMENTATIONS: SUBCLR

Given the difficulty of designing data augmentations for EEG, we sought to construct a method not requiring any augmentations. We assume that the pathology-related statistics we aim to learn from data can be characterised as a source of between-subject information. Whereas Banville et al. (2021) focused on the temporal order of epochs and thereby constructed a proxy-label for self-supervised learning, we rely on the subject identity of samples as a proxy-label. This becomes possible as we do not focus on within-subject tasks such as sleep-staging. This allows for the formulation of a simple contrastive method that does not require augmentations and has a simpler loss objective to interpret. Specifically, we maximize the similarity of embeddings with identical subject identities (i.e. those recorded from the same subject) and minimize the similarity with respect to all samples in a batch with different subject identities. Given the batch embeddings $\mathbf{z}$, we initially follow SimCLR by computing the pairwise cosine similarity:

$$\boldsymbol{z} = \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2} \tag{1}$$

$$\boldsymbol{S} = \frac{\boldsymbol{z}\boldsymbol{z}^\top}{\tau} \tag{2}$$

Next, a mask $M \in \mathbb{R}^{N \times N}$ with batch size N is created using the subject identities $\boldsymbol{p} \in \mathbb{R}^N$:

$$M_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{p}_i = \boldsymbol{p}_j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$M_{ii} = 0, \text{for all } i \in \{1, ..., N\} \tag{4}$$

This allows for the loss computation:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(-\frac{\sum_{j=1}^{N} \mathbf{LSM}_{ij} \mathbf{M}_{ij}}{\sum_{j=1}^{N} \mathbf{M}_{ij}} \right) \tag{5}$$

where $\mathbf{LSM} = \log\left(\text{softmax}(\boldsymbol{S})\right)$ which features a temperature parameter and is used to compute log-probabilities.

As conventional batch sampling is entirely stochastic, there is no guarantee to avoid sampling only a single sample for a given subject, which prevents the similarity computation. SubCLR addresses this limitation by introducing a parameter u that controls the number of unique subject identities per batch (where $N \mod u = 0$ and $u > 1$). The sampling process first selects u random subjects, then samples $\frac{N}{u}$ EEG recordings per subject to maintain a constant batch size N. Empirical evaluation on the TUAB training set showed optimal performance with $u = 8$ and temperature $\tau = 0.02$, with the method being relatively robust to hyperparameter choices (Appendix Figure 4).

3.2.2 MODELS

Architecture We use the same neural network architecture for all analyses unless indicated otherwise. For the encoder model f_θ, we use a residual 1D-convolutional neural network applying variable kernel sizes of 4, 8, and 16 samples in parallel. For the projection head g_ϕ, we used a 2-layer MLP. We provide further details on the used architecture as well as optimization in Appendix 9b and A.3.4 respectively.

Evaluation We evaluate the learned representations $\mathbf{h}$ on the downstream classification task by freezing the encoder's weights and training a linear logistic regression model. Specifically, a representation $\mathbf{h}_c$ is obtained for each of C EEG channels, which are concatenated to yield an epoch-level representation $\mathbf{h}_e$. By predicting the binary labels $\mathbf{y}$ we thus yield epoch-level predictions, which are averaged within-subject to derive subject-level predictions. We perform a grid-search for L2 regularization over nine logarithmically-spaced values between 10^{-8} and 10^{8}.

Baselines We further compare SSL methods to supervised deep learning for which we use the same EEG encoder as during pretraining, handcrafted features describing either the time-series or power spectrum based on frequency bands Gemein et al. (2020); Engemann et al. (2022), as well as a Riemmanian filterbank approach (Sabbagh et al., 2019; 2020). These methods have been shown to be strong baselines and we detail their implementation and tuning in Appendix A.3.5.

4 RESULTS

4.1 HOW DO LEARNED REPRESENTATIONS COMPARE FOR PATHOLOGY DETECTION?

Upon comparing SSL pretraining approaches, predictive performance is in general considerably higher for neurological disorders (TUAB) compared to psychiatric ones (HBN; Figure 1A). Many of the methods exhibit broadly similar performance and show comparable scaling properties as the number of exposed labels is increased. We observe that SubCLR scores much better for the HBN data and ranks first and second for the TUAB subsets. Remaining SSL methods show variability in their rank across subsets. BYOL or ContraWR score well on TUAB but none of the augmentation-based methods perform well for the psychiatric HBN data.

All augmentation-based SSL methods perform highly similar to the baselines methods and therefore, as well as visual clarity, we compare to SubCLR as the best performing SSL method (Figure 1B). We note that while the rank-ordering of models changes between datasets, SubCLR performs well across both datasets. Whereas both SSL and supervised deep learning enabled better prediction than expert-based methods for TUAB, we did not observe this for HBN. This may suggest different levels of complexity of pathology-related features across these two domains.

Besides absolute performance of SSL methods (Figure 1A), we also present differences compared to using random weights (i.e. an untrained encoder model; Figure 1C). This delta more clearly shows how much pretraining improved representations as the untrained model controls for inductive biases present in the CNN architecture. However, we note that this comes with some arbitrariness. While we find that larger residual CNNs outperform previously used smaller CNNs (Appendix Figure 5), this difference is markedly larger when models are untrained (Appendix Figure 5). As the random-weights baseline appears to strongly depend on model architecture choices, SSL can appear less effective compared to the literature due to a more performative baseline. We find that benefits over a randomly initialized CNN are surprisingly modest, especially when the NOR subset is used for pretraining.

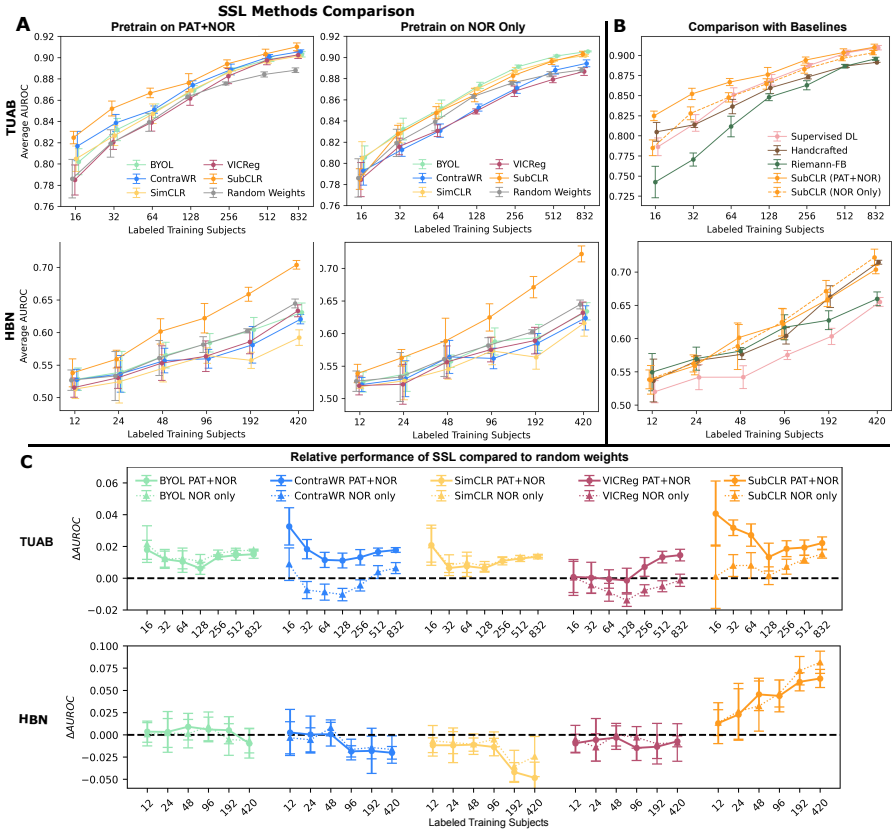

Figure 1: A) A comparison of SSL methods on the TUAB and HBN datasets via averaged AUC scores. Models were trained on a subsample of each dataset consisting of either no pathological subject (NOR) or an equal amount of subjects with and without pathology (PAT+NOR). B) The best performing SSL method is compared against baselines models. C) A different view of results presented in subfigure A. Average scores obtained from using random weights is indicated by the dotted horizontal line. Colored lines indicate average difference scores of SSL methods compared to random weights. Error bars show the standard deviation across cross-validation folds.

Some methods even decrease downstream performance, suggesting inappropriate types of invariance are learned. The used data augmentations were optimized for neurological pathology detection, sleep staging, and emotion recognition by Mohsenvand et al. (2020). This indicates that the benefits of augmentations are dataset and task-specific (Rommel et al., 2022), which we confirm for TUAB and HBN in Appendix Figure 6. This highlights the advantage of SubCLR bypassing the need for searching over augmentations, as it is computationally expensive and may increase the risk of overfitting.

4.2 Do SSL methods benefit from unlabeled pathological samples during pretraining?

We also examined whether methods are able to utilize pathological samples during pretraining. We compare downstream performance following pretraining datasets which were matched for age, sex, and sample size, but either included or excluded pathological subjects (Figure 1A). For TUAB, we find that ContraWR, VICReg, and SubCLR benefit from their inclusion, whereas we observe no such effects for the HBN dataset. This difference is likely related to the more pronounced signatures of neurological pathology for TUAB. Interestingly, a significant portion of SubCLR performance, especially for the HBN data, may be obtained from pretraining on normal samples. One explanation concerns the possibility that psychiatric-related features exhibit considerable variation in less affected populations. These findings may however also indicate that learned representations are of relatively low specificity with respect to pathology.

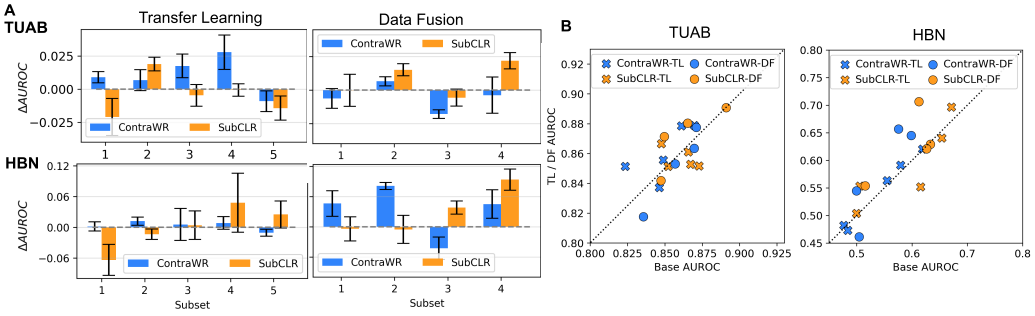

Figure 2: Effects on performance from transfer learning (TL) and data fusion (DF) strategies. A) A visualisation of relative scores of linear evaluation of models pretrained with and without TL or DF. The top and bottom rows use TUAB and HBN as target datasets respectively. Error bars show the standard deviation across cross-validation folds. B) Scatter plot of the absolute scores of models without ('base') and with TL or DF.

4.3 Can learned representations be transferred to small, external data?

Whereas so far we analysed predictive performance within datasets, we also investigated whether external datasets can be used during pretraining to aid downstream pathology classification on small downstream datasets (Figure 2). This is particularly important for the clinical domain, which often features small sample sizes. **Transfer learning:** We evaluated cross-dataset transfer by pretraining on a complete source dataset (either TUAB with n=2711 or HBN with n=2300) and fine-tuning on small subsets (n≈100) of a target dataset. Specifically, these small subsets are obtained by subsampling the training set and creating five age, sex, and pathology matched subsets without overlap (further details in Appendix A.3.1). **Data fusion:** Additionally, we explored combining smaller target data subsets (n≈125 times four without overlap) with a source data subset (n≈500) during pretraining to evaluate potential performance improvements while controlling age and sex distributions.

We observe considerable variation due to simulating small sample sizes, with both significant performance gains and losses. On average, the best performance gains for TUAB are seen when using transfer learning after augmentation-based pretraining on HBN. When HBN is the target, data fusion with TUAB produces better results for both methods. Transfer learning working better for TUAB may indicate that finetuning with limited data is sufficient for neurological but not for psychiatric pathology. The latter domain may rely on more general features of lesser specificity, which can be learned by introducing additional non-psychiatric data via data fusion.

4.4 What are the scaling dynamics of SSL methods with dataset size?

We additionally investigated how the size of the in-distribution pretraining dataset affects downstream performance (Figure 3). We compare downstream performance after pretraining five models on the entire training set (n=2711/2300), the matched subsets (n=928/476), and each of the non-overlapping subsets used to evaluate transfer learning (n=100/94). This analysis also helps understand whether the aforementioned results of transfer learning and data fusion with 'external' data may depend on shifts in data distribution. However, this explanation appears unlikely as we do not observe considerably better downstream performance (while always evaluating at n=100) by increasing pretraining dataset size in the current setting. A minor exception includes TUAB showing more consistent performance gains with SubCLR, albeit of moderate size. In order to investigate whether the poor scaling of models was due to a lack of parameters in the encoder model, we performed the scaling analyses using an encoder model with twice the width and twice the depth (6M instead of 747K parameters). In a further analysis, we had the encoder model operate on time-frequency data instead of the EEG signal directly. Neither of these adjustments materially affected the scaling behaviour (Appendix Figure 7).

To assess whether limited scaling was specific to pathology detection, we evaluated performance on age- and sex-prediction tasks (Appendix Figure 8). SubCLR again performed better but plateaued at

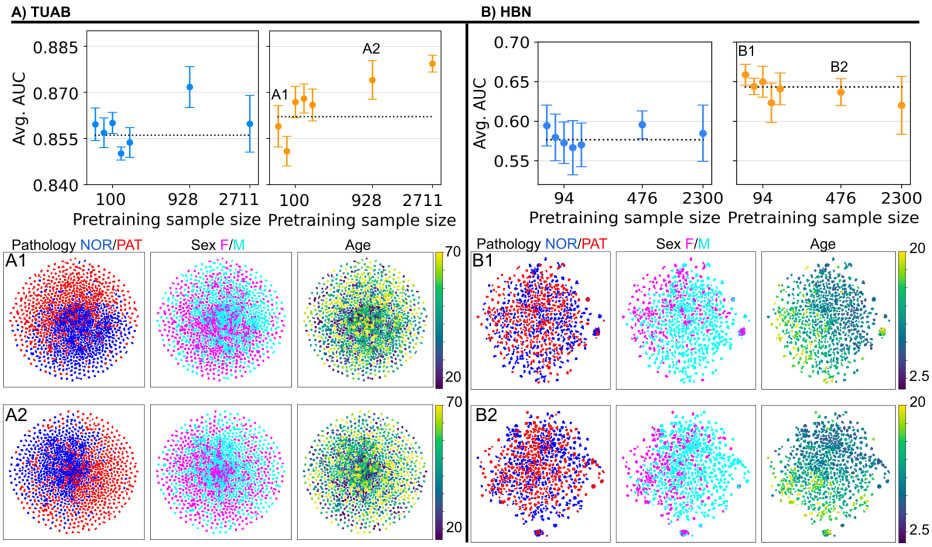

Figure 3: A,B) The effect of increases in pretraining sample sizes on downstream performance is inspected for augmentation-based SSL (blue; ContraWR) and subject-based SSL (SubCLR; orange). Dotted horizontal lines indicate the average AUC across the five data subsets (n=100 under A and n=94 under B). Error bars show the standard deviation across cross-validation folds. We include visualisations of t-SNE projections of learned representations.

100 recordings, while augmentation-based methods showed better scaling but lower overall performance. T-SNE visualizations of learned representations for pathology, age, and sex showed no meaningful progression with increased sample sizes (Figure 3). This early saturation suggests learned representations have limited complexity, aligning with findings from the 46M parameter LaBraM model on TUAB, which gained only $\approx$1.5% performance when increasing pretraining data from 100 to 2500 hours (Jiang et al., 2024).

5 Discussion

We present various insights into the use of SSL for pathology detection with EEG data. First, we show that subject identities can be used to explicitly promote the encoding of between-subject information during SSL, improving pathology detection over augmentation-based methods on two datasets. We furthermore contrast various analyses between the detection of neurological and psychiatric pathology. We observe that only for the neurological domain does SSL outperform all baseline methods and can learn better features by including pathological samples during pretraining. Furthermore, the efficacy of transfer learning and data fusion depended on the target domain. However, we observed limited scaling with respect to pretraining sample size. And although SSL performed well on both datasets, learned representations may be of low complexity and specificity. These findings have important implications for clinical applications and suggest that the clinical domain should be considered.

Some considerations of the current study deserve mention. First, the investigated datasets differed not only in the domain of pathology, but also in other regards including demographics, EEG system, and sample size. These may therefore have contributed to the observed differences between datasets. Although we filtered out pathological cases from data subsets based on labels, it is expected they still contained variable degrees of pathology, likely making our empirical contrast imperfect. While concurrent work explores alternative SSL approaches, our results highlight the importance of developing methods that can effectively scale with data size and capture domain-specific pathological patterns, particularly for psychiatric applications.

MEANINGFULNESS STATEMENT

Neural recordings during rest capture fundamental aspects of human life, including pathological variations in brain function. These deviations from typical function are not merely abnormalities, but represent essential dimensions of life that meaningful representations must capture. Our systematic investigation of how self-supervised learning encodes such information, particularly the distinct signatures of neurological versus psychiatric conditions, advances our understanding of how to learn representations that respect both the universality and diversity of human neural function. This work contributes to the broader goal of developing more comprehensive and nuanced models of human brain activity.

REFERENCES

Obada Al Zoubi, Chung Ki Wong, Rayus T Kuplicki, Hung-wen Yeh, Ahmad Mayeli, Hazem Refai, Martin Paulus, and Jerzy Bodurka. Predicting age from brain eeg signals—a machine learning approach. *Frontiers in aging neuroscience*, 10:184, 2018.

Lindsay M Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data*, 4(1):1–26, 2017.

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 2390–2397. IEEE, 2008.

Mohammad R Arbabshirani, Sergey Plis, Jing Sui, and Vince D Calhoun. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, 145:137–165, 2017.

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.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Colin D Binnie and Hermann Stefan. Modern electroencephalography: its role in epilepsy management. *Clinical Neurophysiology*, 110(10):1671–1697, 1999.

Yun-Hao Cao and Jianxin Wu. A random cnn sees objects: One inductive bias of cnn and its applications. In *Proceedings Of The AAAI Conference On Artificial Intelligence*, volume 36, pp. 194–202, 2022.

Adele E Cave and Robert J Barry. Sex differences in resting eeg in healthy young adults. *International Journal of Psychophysiology*, 161:35–43, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

Mohammad-Javad Darvishi-Bayazi, Mohammad Sajjad Ghaemi, Timothee Lesort, Md Rifat Arefin, Jocelyn Faubert, and Irina Rish. Amplifying pathological detection in eeg signaling pathways through cross-dataset transfer learning. *Computers in Biology and Medicine*, 169:107893, 2024.

Arent de Jongh, Jan Casper de Munck, Sónia I Gonçalves, and Pauly Ossenblok. Differences in meg/eeg epileptic spike yields explained by regional differences in signal-to-noise ratios. *Journal of clinical neurophysiology*, 22(2):153–158, 2005.

Alexandru Dimofte, Glenn Anta Bucagu, Thorir Mar Ingolfsson, Xiaying Wang, Andrea Cossettini, Luca Benini, and Yawei Li. Cerebro: Compact encoder for representations of brain oscillations using efficient alternating attention. *arXiv preprint arXiv:2501.10885*, 2025.

Bruno Dubois, Nicolas Villain, Giovanni B Frisoni, Gil D Rabinovici, Marwan Sabbagh, Stefano Cappa, Alexandre Bejanin, Stéphanie Bombois, Stéphane Epelbaum, Marc Teichmann, et al. Clinical diagnosis of alzheimer's disease: recommendations of the international working group. *The Lancet Neurology*, 20(6):484–496, 2021.

Denis A Engemann, Federico Raimondo, Jean-Rémi King, Benjamin Rohaut, Gilles Louppe, Frédéric Faugeras, Jitka Annen, Helena Cassol, Olivia Gosseries, Diego Fernandez-Slezak, et al. Robust eeg-based cross-site and cross-protocol classification of states of consciousness. *Brain*, 141(11):3179–3192, 2018.

Denis A Engemann, Apolline Mellot, Richard Höchenberger, Hubert Banville, David Sabbagh, Lukas Gemein, Tonio Ball, and Alexandre Gramfort. A reusable benchmark of brain-age prediction from m/eeg resting-state signals. *Neuroimage*, 262:119521, 2022.

Massimo Filippi, Maria A Rocca, Olga Ciccarelli, Nicola De Stefano, Nikos Evangelou, Ludwig Kappos, Alex Rovira, Jaume Sastre-Garriga, Mar Tintorè, Jette L Frederiksen, et al. Mri criteria for the diagnosis of multiple sclerosis: Magnims consensus guidelines. *The Lancet Neurology*, 15 (3):292–303, 2016.

Claas Flint, Micah Cearns, Nils Opel, Ronny Redlich, David MA Mehler, Daniel Emden, Nils R Winter, Ramona Leenings, Simon B Eickhoff, Tilo Kircher, et al. Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology*, 46(8):1510–1517, 2021.

Robert Freedman, David A Lewis, Robert Michels, Daniel S Pine, Susan K Schultz, Carol A Tamminga, Glen O Gabbard, Susan Shur-Fen Gau, Daniel C Javitt, Maria A Oquendo, et al. The initial field trials of dsm-5: new blooms and old thorns, 2013.

Ahmed El Gazzar, Rajat Mani Thomas, and Guido Van Wingen. Improving the diagnosis of psychiatric disorders with self-supervised graph state space models. *arXiv preprint arXiv:2206.03331*, 2022.

Lukas AW Gemein, Robin T Schirrmeister, Patryk Chrabaszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of eeg pathology. *NeuroImage*, 220:117021, 2020.

Sam Gijsen and Kerstin Ritter. Eeg-language modeling for pathology detection. *arXiv preprint arXiv:2409.07480*, 2024.

Daniel M Goldenholz, Seppo P Ahlfors, Matti S Hämäläinen, Dahlia Sharon, Mamiko Ishitobi, Lucia M Vaina, and Steven M Stufflebeam. Mapping the signal-to-noise-ratios of cortical sources in magnetoencephalography and electroencephalography. *Human brain mapping*, 30(4):1077–1086, 2009.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Howard H Feldman, Giovanni B Frisoni, Harald Hampel, William J Jagust, Keith A Johnson, David S Knopman, et al. A/t/n: an unbiased descriptive classification scheme for alzheimer disease biomarkers. *Neurology*, 87(5):539–547, 2016.

Mainak Jas, Denis A Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage*, 159:417–429, 2017.

Mainak Jas, Eric Larson, Denis A Engemann, Jaakko Leppäkangas, Samu Taulu, Matti Hämäläinen, and Alexandre Gramfort. A reproducible meg/eeg group study with the mne software: recommendations, quality assessments, and good practices. *Frontiers in neuroscience*, 12:530, 2018.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. https://openreview.net/forum?id=QzTpTRVtrP, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP. OpenReview.

Jin Jing, Aline Herlopian, Ioannis Karakis, Marcus Ng, Jonathan J Halford, Alice Lam, Douglas Maus, Fonda Chan, Marjan Dolatshahi, Carlos F Muniz, et al. Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms. *JAMA neurology*, 77(1): 49–57, 2020.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Joseph Kambeitz, Carlos Cabral, Matthew D Sacchet, Ian H Gotlib, Roland Zahn, Mauricio H Serpa, Martin Walter, Peter Falkai, and Nikolaos Koutsouleris. Reply to: sample size, model robustness, and classification accuracy in diagnostic multivariate neuroimaging analyses. *Biological psychiatry*, 84(11):e83–e84, 2018.

Hassan Aqeel Khan, Rahat Ul Ain, Awais Mehmood Kamboh, Hammad Tanveer Butt, Saima Shafait, Wasim Alamgir, Didier Stricker, and Faisal Shafait. The nmt scalp eeg dataset: an open-source annotated dataset of healthy and pathological eeg recordings for predictive modeling. *Frontiers in neuroscience*, 15:755817, 2022.

Mariam Khayretdinova, Alexey Shovkun, Vladislav Degtyarev, Andrey Kiryasov, Polina Pshonkovskaya, and Ilya Zakharov. Predicting age from resting-state scalp eeg signals with deep convolutional neural networks on td-brain dataset. *Frontiers in Aging Neuroscience*, 14:1367, 2022.

Ann-Kathrin Kiessner, Robin T Schirrmeister, Joschka Boedecker, and Tonio Ball. Reaching the ceiling? empirical scaling behaviour for deep eeg pathology classification. *Computers in Biology and Medicine*, pp. 108681, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Marcus W Koch, Jop Mostert, Pavle Repovic, James D Bowen, Bernard Uitdehaag, and Gary Cutter. Reliability of outcome measures in clinical trials in secondary progressive multiple sclerosis. *Neurology*, 96(1):e111–e120, 2021.

Nicolas Langer, Martyna Beata Plomecka, Marius Tröndle, Anuja Negi, Tzvetan Popov, Michael Milham, and Stefan Haufe. A benchmark for prediction of psychiatric multimorbidity from resting eeg data in a large pediatric sample. *NeuroImage*, 258:119348, 2022.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Sandra K Loo, James J McGough, James T McCracken, and Susan L Smalley. Parsing heterogeneity in attention-deficit hyperactivity disorder using eeg-based subgroups. *Journal of Child Psychology and Psychiatry*, 59(3):223–231, 2018.

Raman K Malhotra and Alon Y Avidan. Sleep stages and scoring technique. *Atlas of sleep medicine*, pp. 77–99, 2013.

Scott Marek and Timothy O Laumann. Replicability and generalizability in population psychiatric neuroimaging. *Neuropsychopharmacology*, 50(1):52–57, 2024.

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

DV Moretti, M Pievani, C Geroldi, G Binetti, O Zanetti, PM Rossini, and GB Frisoni. Eeg markers discriminate among different subgroup of patients with mild cognitive impairment. *American Journal of Alzheimer's Disease & Other Dementias®*, 25(1):58–73, 2010.

Johannes Müller-Gerking, Gert Pfurtscheller, and Henrik Flyvbjerg. Designing optimal spatial filters for single-trial eeg classification in a movement task. *Clinical neurophysiology*, 110(5):787–798, 1999.

Emily Neuhaus, Sarah J Lowry, Megha Santhosh, Anna Kresse, Laura A Edwards, Jack Keller, Erin J Libsack, Veronica Y Kang, Adam Naples, Allison Jack, et al. Resting state eeg in youth with asd: age, sex, and relation to phenotype. *Journal of neurodevelopmental disorders*, 13:1–15, 2021.

Jennifer J Newson and Tara C Thiagarajan. Eeg frequency bands in psychiatric disorders: a review of resting state studies. *Frontiers in human neuroscience*, 12:521, 2019.

NICE. Overview. dementia: Assessment, management and support for people living with dementia and their carers. guidance. nice. 2018.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maria Carla Piastra, Andreas Nüßing, Johannes Vorwerk, Maureen Clerc, Christian Engwer, and Carsten H Wolters. A comprehensive study on electroencephalography and magnetoencephalography sensitivity to cortical and subcortical sources. *Human Brain Mapping*, 42(4):978–992, 2021.

Rebecca B Price, Kathleen Gates, Thomas E Kraynak, Michael E Thase, and Greg J Siegle. Data-driven subgroups in depression derived from directed functional connectivity paths at rest. *Neuropsychopharmacology*, 42(13):2623–2632, 2017.

Geoffrey M Reed, Pratap Sharan, Tahilia J Rebello, Jared W Keeley, María Elena Medina-Mora, Oye Gureje, José Luis Ayuso-Mateos, Shigenobu Kanba, Brigitte Khoury, Cary S Kogan, et al. The icd-11 developmental field study of reliability of diagnoses of high-burden mental disorders: results among adult patients in mental health settings of 13 countries. *World psychiatry*, 17(2): 174–186, 2018.

Cédric Rommel, Joseph Paillard, Thomas Moreau, and Alexandre Gramfort. Data augmentation for learning predictive models on eeg: a systematic comparison. *Journal of Neural Engineering*, 19 (6):066020, 2022.

Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. Chrononet: A deep recurrent neural network for abnormal eeg identification. In *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*, pp. 47–56. Springer, 2019.

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A Engemann. Manifold-regression to predict from meg/eeg brain signals without source modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A Engemann. Predictive regression modeling with meg/eeg: from source power to signals and cognitive states. *NeuroImage*, 222:116893, 2020.

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.

David Shaffer, Madelyn S Gould, James Brasic, Paul Ambrosini, Prudence Fisher, Hector Bird, and Satwant Aluwahlia. A children's global assessment scale (cgas). *Archives of General psychiatry*, 40(11):1228–1231, 1983.

Markos G Tsipouras. Spectral information of eeg signals with respect to epilepsy classification. *EURASIP Journal on Advances in Signal Processing*, 2019(1):1–17, 2019.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

Michel JAM Van Putten, Taco Kind, Frank Visser, and Vera Lagerburg. Detecting temporal lobe seizures from scalp eeg recordings: a comparison of various features. *Clinical neurophysiology*, 116(10):2480–2489, 2005.

Michel JAM Van Putten, Sebastian Olbrich, and Martijn Arns. Predicting sex from brain rhythms with deep learning. *Scientific reports*, 8(1):3069, 2018.

Ziwei Wang and Paolo Mengoni. Seizure classification with selected frequency bands and eeg montages: a natural language processing approach. *Brain Informatics*, 9(1):11, 2022.

Devon Watts, Rafaela Fernandes Pulice, Jim Reilly, Andre R Brunoni, Flávio Kapczinski, and Ives Cavalcante Passos. Predicting treatment response using eeg in major depressive disorder: A machine-learning meta-analysis. *Translational psychiatry*, 12(1):332, 2022.

David Western, Timothy Weber, Rohan Kandasamy, Felix May, Samantha Taylor, Yixuan Zhu, and Luke Canham. Automatic report-based labelling of clinical eegs for classifier training. In *2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–6. IEEE, 2021.

Chaoqi Yang, Danica Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised eeg representation learning for automatic sleep staging. *arXiv preprint arXiv:2110.15278*, 2021.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36, 2024.

Chia-Yen Yang, Pin-Chen Chen, and Wen-Chen Huang. Cross-domain transfer of eeg to eeg or ecg learning for cnn classification models. *Sensors*, 23(5):2458, 2023.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Kaishuo Zhang, Neethu Robinson, Seong-Whan Lee, and Cuntai Guan. Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network. *Neural Networks*, 136:1–10, 2021.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022a.

Zhi Zhang, Sheng-hua Zhong, and Yan Liu. Ganser: A self-supervised data augmentation framework for eeg-based emotion recognition. *IEEE Transactions on Affective Computing*, 2022b.

A Appendix

A.1 Additional Analyses

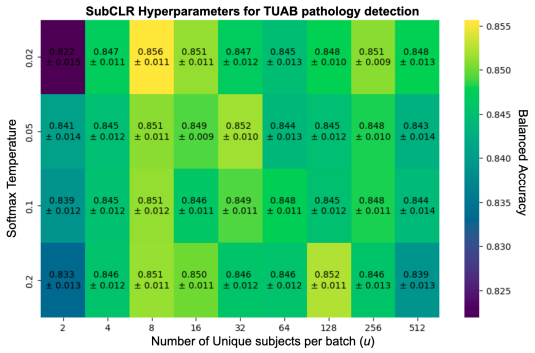

Figure 4: An exploration of the sensitivity of SubCLR to hyperparameters. These include the temperature parameter used in the softmax operation of the loss computation as well as the amount of unique subjects sampled for each batch (u). Pretraining as well as linear evaluation of pretrained models is performed on subsets of the TUAB training set without test set overlap. We find relative insensitivity to the hyperparameters. Even extremes of u (two or 512 different subjects are included in each batch) only result in minor performance degradation. In our study, we use $u = 8$ and a temperature of 0.02.

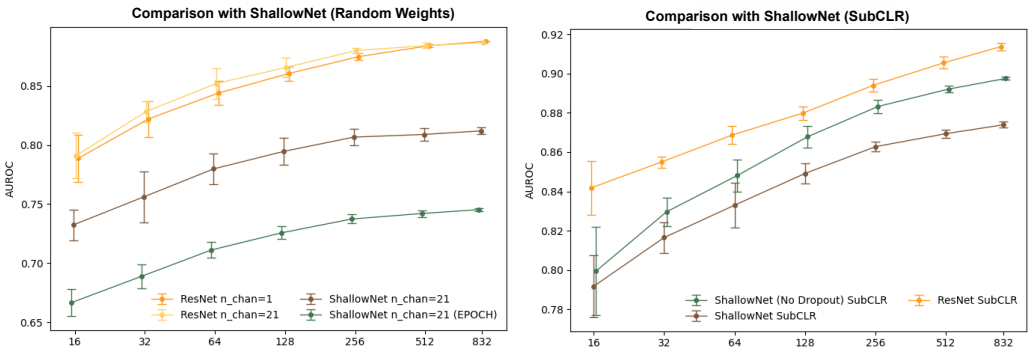

Figure 5: The ShallowNet model introduced by Schirrmeister et al. (2017) is used in Banville et al. (2021) for pathology detection on the TUAB dataset. Besides using the model for SSL, it was also used as a baseline comparison with randomly initialized weights. Left: ShallowNet with random weights was compared against the untrained residual network used in the present work. The set-up as described in Banville et al. (2021) is presented in green, which operates on all 21 EEG channels and is evaluated on an epoch level. (Using balanced accuracy, performance plateaus around 67%, which is similar to the authors' presented results.) The current study, however, performs subject-level inference by averaging model predictions across epochs in a within-subject fashion, which is shown for ShallowNet in brown. The residual network shows considerably better performance without training, both when operating on all 21 EEG channels or when operating on a single-channel basis (yellow and orange respectively). Right: We pretrain ShallowNet using the SubCLR method (brown) and compare it against the residual model (orange). A clear difference in predictive performance is observed. We further show that disabling dropout in ShallowNet during pretraining (green) reduces the delta in performance.

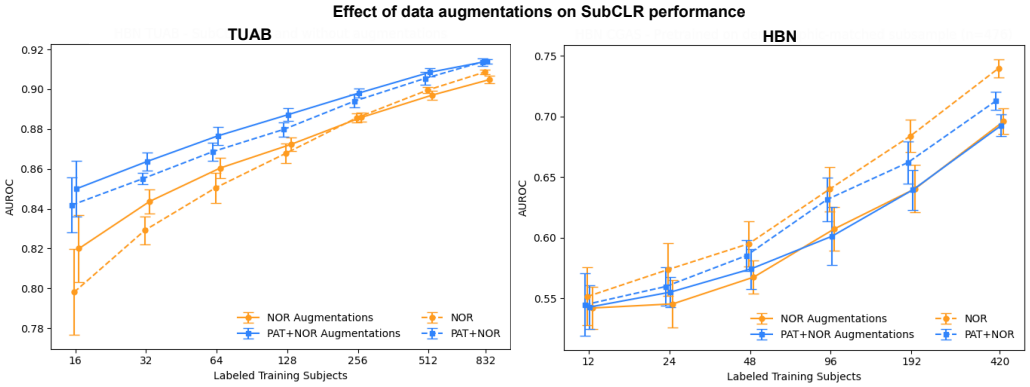

Figure 6: The effect of data augmentations when using SubCLR as a pretraining objective. For TUAB (left), except for larger data regimes the inclusion of data augmentations increases performance, whether the NOR or PAT+NOR data subset is used. Meanwhile, for HBN (right) including data augmentations hurts performance for both data subsets and all number of labeled subjects.

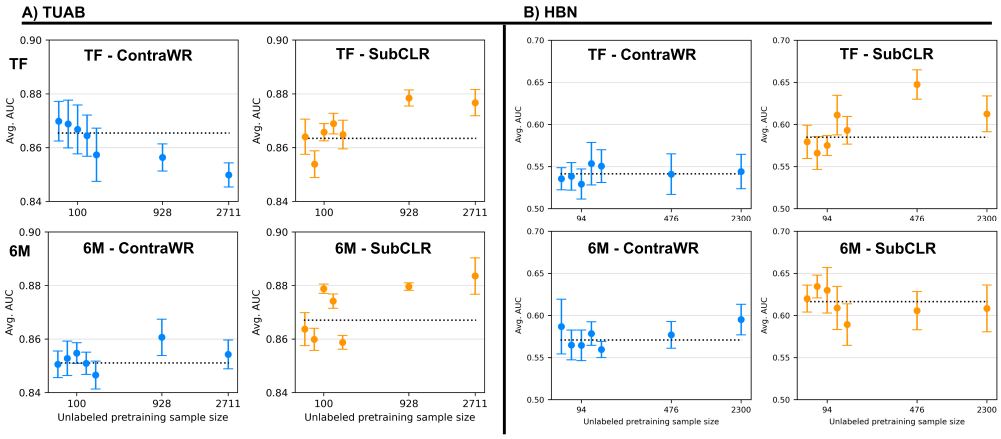

Figure 7: The analyses of the scaling properties with respect to pretraining sample sizes are replicated for the following two model adjustments. First, the input data undergoes a time-frequency decomposition into three bands ([1-7Hz, 8-30Hz, 31-49Hz]), denoted 'TF'. For the ContraWR implementation, we only sample from the Gaussian noise, time-shift, and bandstop-filter data augmentations and apply these to the EEG signal prior to the decomposition. The encoder model thus now operates on three input channels instead of one. Second, the encoder model is made twice as deep and wide, increasing the amount of parameters from 747K to 6M (thus denoted '6M'). Due to the increased computational demands these models were pretrained only once, with error bars therefore reflecting only the standard deviation across K-Fold cross-validation of the linear evaluation (5 repetitions with $K = 5, n = 100$). Results suggest no clear violation of the main results, with significant pretraining sample size increases of up to 23x and 27x yielding very minor or even no classification improvements. An exception is observed for SubCLR-TF on HBN, where an improvement is observed of 0.06 AUC at $n = 476$. However, this improvement merely matches the performance of the base SubCLR model at $n = 94$ (Figure 3) and does not persist at $n = 2300$, indicating it may be due to randomness in pretraining. Our findings appear therefore not dependent on the choice of input data or a lack of parameters in the encoder model. Horizontal, dotted lines indicate the average score across the five smallest subsamples ($n = 94$ or 100).

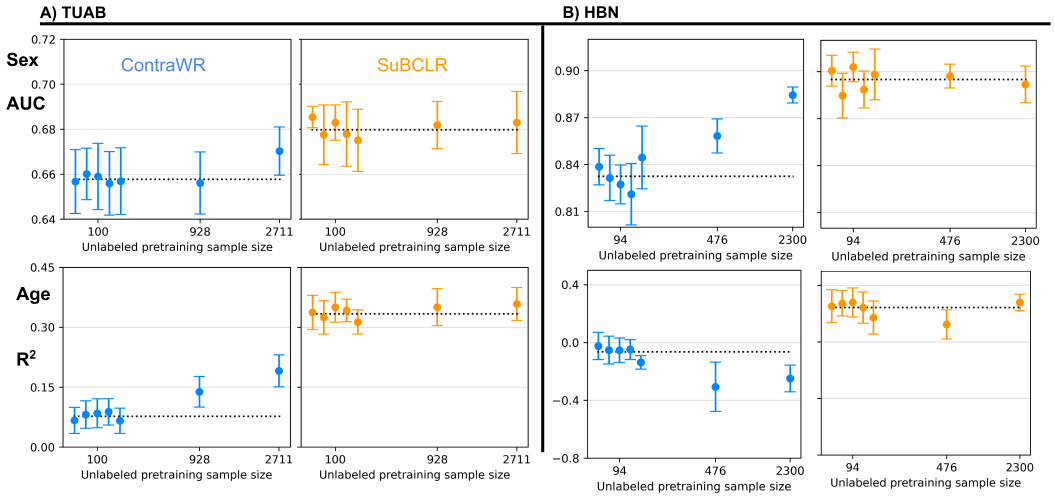

Figure 8: The scaling analysis of pretraining sample size was repeated for downstream age and sex prediction. Linear evaluation is performed using 100 labeled EEG recordings (error bars indicate standard deviations of 5 repetitions of K-Fold cross-validation with $K = 5$). We find for both targets on both datasets of TUAB (A) and HBN (B) higher performance for SubCLR (orange) than ContraWR (blue). We further observe that while SubCLR performance on these tasks has saturated at 100 EEG recordings, ContraWR scales with pretraining sample size except for age prediction on the HBN dataset. However, it remains uncertain whether ContraWR can match or exceed SubCLR with additional data.

A.2 Extended Discussion

The explicit encoding of between-subject information via SubCLR is found to yield representations better suited for pathology detection than augmentation-based SSL on both datasets across all investigated sample sizes. This may result from SubCLR better encoding pathology-related information per se. Alternatively, features related to non-pathological, between-subject factors (e.g. demographics) may be better learned and could be informative for pathology. That is, insofar as data dimensions relating to pathology and other between-subject factors overlap, encoding any such information via SubCLR may have aided downstream pathology detection. For example, oscillatory activity in frequency bands of up to 30 Hz (i.e., delta, theta, alpha, and beta bands) has been informative to predict both age (Al Zoubi et al., 2018; Neuhaus et al., 2021; Khayretdinova et al., 2022) and sex (Van Putten et al., 2018; Cave & Barry, 2021; Neuhaus et al., 2021). These features have also been correlated with pathology detection on TUAB (Gemein et al., 2020) and specific pathologies present in the corpus such as epilepsy (Tsipouras, 2019; Wang & Mengoni, 2022). Moreover, such evidence is also reported for pediatric and psychiatric pathology as found in the HBN cohort (Newson & Thiagarajan, 2019; Neuhaus et al., 2021). In line with this possibility, additional analyses indicated that features learned via SubCLR enabled better age and sex prediction, although it remains unknown to what extent this contributed to the better pathology detection. Meanwhile, as augmentation-based methods were often not able to improve pathology detection considerably over a randomly initialized CNN, their learned features appear to be unspecific to the pathology detection task.

A further finding concerns the similar performance of the augmentation-based SSL methods within each dataset. Yang et al. (2021), in their introduction of ContraWR for sleep staging with EEG data, provide a comparison with BYOL and SimCLR, among others. Although ContraWR performed best, differences between models appeared small, with the rank-order of models changing across the number of exposed labels, potentially indicating considerable noise. While similar performance may be regarded as unsurprising given that these methods all promote augmentation-invariance in the encoder model, it is not consistently found in the computer vision literature. Indeed, BYOL and VICReg were developed in response to SimCLR (Chen et al., 2020) and can show notably better performance, although this is dependent on the specific dataset (Grill et al., 2020; Bardes et al., 2021). This is commonly attributed to these methods being less prone to suffer informational col-

lapse, where representations span a space of reduced dimensionality (Bardes et al., 2021; Jing et al., 2021). The homogeneity of performance observed in the present work may indicate that differing levels of informational collapse did not meaningfully affect results. This is likely when models learn predominantly lower-order, surface statistics of data and thus do not require the efficient and maximal use of the available representational capacity. Compared to natural images, neuroimaging data such as EEG has a markedly lower signal-to-noise ratio (de Jongh et al., 2005; Goldenholz et al., 2009; Piastra et al., 2021). This reduces the ability to accurately compare similar models per se, potentially contributing to the observations made here and the results of Yang et al. (2021). Second, noise likely also reduces the complexity of learnable data statistics. The need for high dimensional representations may thereby be reduced which, in turn, could lessen the impact of informational collapse, reducing the variability between methods.

The interpretation that learned representations are of low-complexity fits with the remarkable performance of using an untrained encoder. As we find that gaps in performance between the currently employed residual CNN and a smaller CNN (ShallowNet by (Schirrmeister et al., 2017)) considerably shrink following pretraining, the inherent inductive biases of CNNs may account for a considerable portion of performance. These CNNs share some such biases, including local connectivity (adjacent time points are more related than distance ones), translation invariance (features are independent of their position in time), and a non-linear mapping between input and output (LeCun et al., 1995). Meanwhile, inductive biases also depend on specific aspects of model architecture: the additional convolutional layers of the residual CNN enable feature learning across different temporal scales and the iterative pooling operations assume different levels of informational redundancy. The observation that randomly initialized CNNs are powerful models is widely recognized in the machine learning literature (Ulyanov et al., 2018), inspired early version of the BYOL approach (Grill et al., 2020), and was found to scale with network depth (Cao & Wu, 2022), which is compatible with our observations. Yet, while for computer vision augmentation-based pretraining may add a minimum of 10% absolute accuracy (Cao & Wu, 2022), we observe considerably lower gains.

Whereas positive results for transfer learning have been reported for fully-supervised learning in domains such as sleep staging (Yang et al., 2023) and motor imagery classification (Zhang et al., 2021), its benefit for pathology remains uncertain. Recent work investigated transfer learning and data fusion for pathology detection on TUAB and an additional pathology dataset (Darvishi-Bayazi et al., 2024). TUAB was used as the source dataset for transfer learning, where only one of four investigated models showed an improvement by being pretrained on TUAB. Data fusion analyses worked better using TUAB as a source, improving classification for three out of four cases. Meanwhile, data fusion using TUAB as a target dataset lead to performance decreases for all four models.

On possible low complexity of learned representations and scaling limitations

While similar performance across augmentation-invariant methods might seem expected, it contrasts with computer vision, where BYOL and VICReg often outpace SimCLR (Chen et al., 2020), although this is dataset dependent (Grill et al., 2020; Bardes et al., 2021). This is commonly attributed to these methods being less prone to suffer informational collapse, where representations span a space of reduced dimensionality (Bardes et al., 2021; Jing et al., 2021). The homogeneity of performance observed in the present work may indicate that differing levels of informational collapse did not meaningfully affect results. This is likely when models learn predominantly lower-order, surface statistics of data and thus do not require the efficient and maximal use of the available representational capacity. Compared to natural images, neuroimaging data such as EEG has a markedly lower signal-to-noise ratio (de Jongh et al., 2005; Goldenholz et al., 2009; Piastra et al., 2021). This impairs model differentiation and simplifies learnable features, reducing the need for high-dimensional representations and muting collapse effects.

The interpretation that learned representations are of low-complexity fits with the remarkable performance of using an untrained encoder. Pretraining shrinks gaps between our residual CNN and ShallowNet (Schirrmeister et al., 2017)), hinting CNN biases drive a considerable portion of the result. These biases vary by design: extra layers in the residual CNN capture multi-scale temporal features, while pooling assumes redundancy. The observation that randomly initialized CNNs are powerful models is widely recognized in the machine learning literature (Ulyanov et al., 2018), inspired early version of the BYOL approach (Grill et al., 2020), and was found to scale with network depth (Cao & Wu, 2022), which is compatible with our observations. Yet, while for computer vision

augmentation-based pretraining may add a minimum of 10% absolute accuracy (Cao & Wu, 2022), we observe considerably lower gains.

The performance plateau with more pretraining data might reflect multiple issues. Architectural constraints, like our CNN's focus on local patterns, could saturate early, which is not alleviated by scaling up the encoder model to 6M parameters. Furthermore, single-channel pretraining naturally limits the complexity of learned representations as spatial relationships are not present in the input. EEG signal limitations, high signal-to-noise, likely place a ceiling on complexity too, also seen in LaBraM's modest scaling (Jiang et al., 2024). We propose that alternative pretraining strategies may offer a solution when they promote the learning of more finegrained information which is not restricted to mere subject-level differences or based on difficult-to-formulate EEG augmentations.

Novelty compared to related work

In this work, we advance the application of self-supervised learning (SSL) to clinical EEG data by introducing SubCLR, a novel contrastive learning approach that explicitly encodes between-subject information using subject identity as a proxy, bypassing the need for data augmentations—a key departure from prior augmentation-based SSL methods like SimCLR, BYOL, and VICReg (e.g., (Chen et al., 2020; Grill et al., 2020; Bardes et al., 2021)). Unlike earlier work that often relied on unmatched datasets, architectures, and limited baseline comparisons, we conduct a systematic evaluation on two large clinical EEG datasets (TUAB and HBN), controlling for demographic confounds and benchmarking multiple SSL strategies against supervised deep learning and handcrafted feature baselines. Our approach uniquely achieves superior pathology detection across both neurological and psychiatric domains, revealing distinct representation learning dynamics—such as SubCLR's robustness to limited labeled data and its ability to leverage unlabeled pathological samples, particularly in neurological contexts. Additionally, we explore scaling dynamics with dataset size and transferability to small external datasets, aspects underexplored in prior EEG SSL research, providing new insights into the practical utility and limitations of SSL for clinical applications. These contributions collectively offer a more nuanced understanding of representation learning for EEG, tailored to the heterogeneity of clinical pathology, beyond what has been previously demonstrated.

A.3 Extended Methods

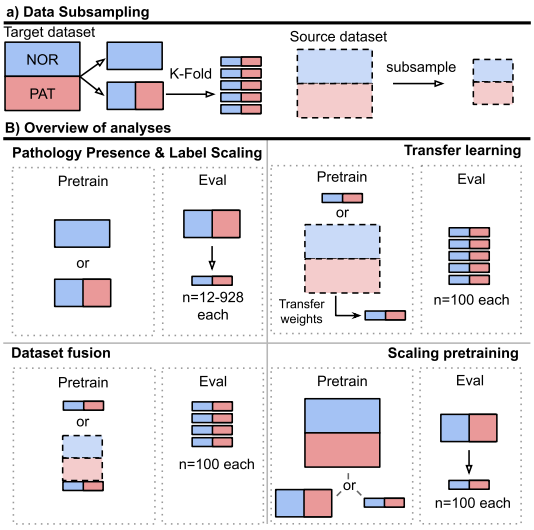

Figure 9a. The carried out analyses involve multiple dataset subsampling steps to allow a comparison of performance between differing pretraining setups. In A), the dataset subsampling is visualized. Meanwhile, B) includes an overview of the undertaken analyses, which involve varying conditions of label-free pretraining and an appropriate evaluation setup.

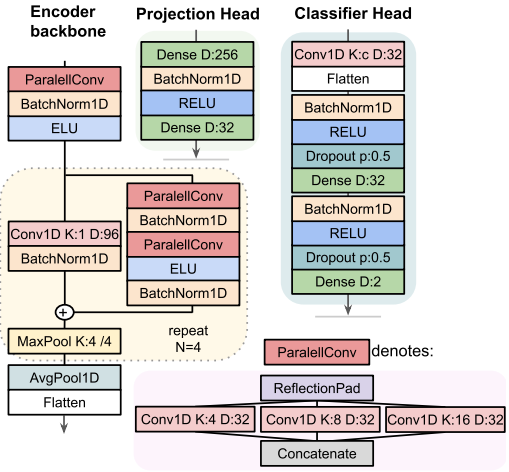

Figure 9b. A visual depiction of the employed CNN architectures. The encoder backbone is used for all SSL pretraining analyses as well as supervised learning comparisons. The encoder is combined with the projection head during pretraining and discarded afterwards, while the classifier head is used for supervised learning. D: Output dimensionality, K: Kernel size, c: Number of EEG channels

A.3.1 DATA PREPROCESSING AND SUBSAMPLING

Subsampling From the TUAB training set we create subsets containing only normal recordings (NOR) or an equal amount of abnormal and normal recordings (NOR+PAT). We obtain a female ratio of 0.5 and match the age-distributions, both between the abnormal and normal recordings in the NOR+PAT subset as well as between the NOR and NOR+PAT subsets themselves, yielding $n = 928$ for each subset (Table 1). For the HBN dataset, we construct half of the NOR+PAT subset by including all subjects with a CGAS score of 50 or lower, resulting in $n = 238$. Next, the NOR subset is created by including the $n * 2 = 476$ subjects with the highest CGAS scores (73 or higher). Finally, we complete the NOR+PAT subset by sampling 238 subjects from the NOR subset by matching sex and age distributions. We note, however, that this methodology leads to the NOR and NOR+PAT subsets of HBN to differ in sex ratios (59.2% and 66.8% male respectively). Avoiding this entirely would have further shrunk available sample sizes considerably, while our priority was to prevent models trained on NOR+PAT subsets exploiting sex or age difference between normal and pathological data. From the HBN hold-out test set, we similarly only use the extreme groups ($n = 138$). We provide a visualization of the subsampling process in Figure 9a.

Preprocessing. The TUAB EEG recordings feature different amounts of electrodes. As is common in the literature, we use the 21 electrodes shared across subjects. To reduce the impact of differences in recording lengths between subjects, we use a maximum of 11 minutes per recording. For the HBN, we follow (Langer et al., 2022) by discarding those channels mostly recording muscular activity by being located on the chin and neck, including: E1, E8, E14, E17, E21, E25, E32, E48, E49, E56, E63, E68, E73, E81, E88, E94, E99, E107, E113, E119, E125, E126, E127, and E128, resulting in 104 electrodes. The resting-state recording included 'eyes-closed' and 'eyes-open' segments, which we included both in order to maximize the amount of data. Both datasets were bandpass filtered to 0.1 - 49Hz, the data was resampled to 200Hz, and split into epochs with a length of 10 seconds. Due to the lower amount of channels for TUAB, epochs were rejected using the 'Global' setting of AutoReject (Jas et al., 2017), which sets a singular peak-to-peak rejection threshold for all channels. As the HBN dataset featured more channels and we aimed to preserve data given potential extra noise due to the pediatric nature, the 'Local' setting was used, setting a threshold on a per channel basis. The default maximum number of channels to be interpolated was increased to accommodate a channel count of 104, performing cross-validation over [4, 16, 46] channels. Finally, both datasets were re-referenced with an average reference. The Python-based software MNE and MNE-BIDS-Pipeline were used for preprocessing (Jas et al., 2018).

A.3.2 AUGMENTATION-BASED SSL

Given the similarity between methods, we briefly describe the SimCLR method by Chen et al. (2020) and then denote the important differences for subsequent methods. Given $\mathbf{x}$, two batches of different views $\tilde{\mathbf{x}}_\mathbf{m}$ and $\tilde{\mathbf{x}}_\mathbf{n}$ are created by applying data augmentations. These are passed through an encoder model f_θ creating the representations $\mathbf{h}_m$ and $\mathbf{h}_n$, which in turn are passed through a projection head g_ϕ to produce embeddings $\mathbf{z}_m$ and $\mathbf{z}_n$. At this point, the loss function is applied which for SimCLR is the InfoNCE contrastive loss (Oord et al., 2018) combined with the cosine similarity metric. The set $\{\tilde{\mathbf{x}}_k\}$ includes a positive pair of samples $\tilde{\mathbf{x}}_m$ and $\tilde{\mathbf{x}}_n$ and the pretraining task may be formulated as identifying $\tilde{\mathbf{x}}_n$ in $\{\tilde{\mathbf{x}}_k\}_{k \neq m}$ given a $\tilde{\mathbf{x}}_m$. In other words, similarity of embedding pairs $(\mathbf{z}_m, \mathbf{z}_n)$ is increased, while their similarity with every other embedding in the batch is decreased. Note that the projector g_ϕ is only used during pretraining and is discarded for any downstream task, where only the encoder f_θ output is used. This follows from empirical observations (Chen et al., 2020), possibly due to later layers being too narrowly optimized for the contrastive loss. Note that trivial solutions in form of identical embeddings for every input are avoided by the negative part of the contrastive loss, which pushes away embeddings from each other.

Later methods attempt to resolve specific issues of SimCLR, which includes a requirement of large batch-sizes as well as dimensional collapse, in which embedding vectors span a lower-dimensional space, hurting performance (Jing et al., 2021). Variance-Invariance-Covariance Regularization (VICReg) aims to avoid collapse explicitly by introducing two regularization terms applied to each of the two sets of embeddings separately (Bardes et al., 2021). First, a 'variance' term ensures the batch-wise standard deviation of each variable of an embedding remains above a threshold, which prevents shrinkage of the embeddings to zero, leading to a trivial solution. Second, a 'covariance' term shrinks the batch-wise covariance between every pair of embedding variables, which decorre-

lates them and prevents dimensional collapse. As a consequence, the contrastive loss is no longer needed, which is replaced by the mean square distance between embedding vectors, which when minimized also promotes similarity of the pair $\tilde{\mathbf{x}}_\mathbf{m}$ and $\tilde{\mathbf{x}}_\mathbf{n}$.

Bootstrap-Your-Own-Latent (BYOL; Grill et al. (2020)) takes an alternate approach to avoid collapse and uses two neural networks, referred to as the online and target networks. Both networks still use an encoder and projection model, with the online encoder model being used for downstream tasks. The main idea is that the online network attempts to predict the output of the target model. Collapse is prevented by updating the weights of the target network by using an exponential moving average of the weights of the online network. As no contrastive loss is required, a mean square distance of ℓ_2 normalized embedding vectors is used. The exponential moving average parameter is set to the recommended 0.996.

Contrast with the World Representation (ContraWR) is an SSL method that was specifically proposed to improve SSL for EEG and evaluated for automatic sleep staging (Yang et al., 2021). It is a contrastive method which uses an aggregated representation of the batch as the negative part. We use their "ContraWR+" method which uses a weighted average to create the aggregated representation, with weights being positively scaled by the relative embedding similarities, effectively creating a harder pretraining task by increasing the influence of samples which are hard to distinguish. Loss-function hyperparameters σ, τ, δ are set to their recommended default values.

A.3.3 AUGMENTATIONS

Here we adopt the augmentations proposed by Mohsenvand et al. (2020) (see Table 2), who showed the benefit of these augmentations for pathology detection on TUAB as well as other tasks. As we apply z-score normalization we reduced the range of the proposed DC shift augmentation.

Data Augmentation	**Min**	**Max**
Amplitude Scale	0.5	2
Time Shift in samples	-50	50
DC shift in mmV	-4	4
Zero-Masking in samples	0	150
Additive Gaussian Noise (σ)	0	0.2
Band-Stop Filter (5Hz width)	2.8	47

Table 2: Data augmentations adapted from Mohsenvand et al. (2020).

A.3.4 ARCHITECTURE AND OPTIMIZATION

A visualization of the EEG encoder architecture is shown in Figure 9b and allows for a flexibility where the amount of residual blocks and number of filters can be easily adapted. We used 4 residual blocks with each block followed by max pooling with a kernel size and stride of 4, leading to incremental downsampling. The number of filters per kernel size was set to 32, thus the final average pooling operation yields a 32×3=96 dimensional representation vector $\mathbf{h}$. This yields a total of 747K trainable parameters in the encoder. For the projector, we used a 2-layer MLP with a hidden dimension of width 256 and an output dimension of 32. For VICReg, the output dimension is recommended to be wider and was thus set to 256 (Bardes et al., 2021).

For optimization, we follow recommendations for large-batch SSL pretraining (Grill et al., 2020). The LARS optimizer is used in combination with a cosine decay learning rate schedule of 50 epochs with a linear warm-up period of 5 epochs (You et al., 2017). As HBN data was recorded with considerably more electrodes, resulting in more highly correlated samples, models were pretrained for 25 epochs with 3 warm-up epochs. The base learning rate is set to 0.3, scaled with the batch size (LearningRate = 0.2 $\times$ BatchSize/256). We use weight decay of 0.0001 while excluding batch normalization and bias parameters from weight decay and LARS adaptation. As methods use different amounts of memory, batch size was set to 2048 for SimCLR, ContraWR, and SubCLR and 1024 for BOYL and VICReg. Either an Nvidia GeForce RTX 3090 or a Tesla V100 GPU was used.

A.3.5 BASELINE MODELS

Supervised deep learning. For supervised learning we employed the same encoder model as for the SSL approaches, but instead of the projector head used a classification head. This consisted of a spatial convolution with 32 filters which was applied across all EEG channels after which output was flattened and passed through a 3-layer MLP with dropout ($p = 0.5$). We use the Adam optimizer with a weight-decay of 0.001, and reduce the learning rate by half when validation loss does not improve for three consecutive epochs, with early stopping after six consecutive epochs without improvement (Kingma & Ba, 2014). For TUAB, we use a batch size of 180 and a learning rate of 0.0003, while the larger number of EEG sensors for HBN necessitates a smaller batch size of 36 and a lower learning rate of 0.0001.

Handcrafted features. Machine learning using EEG data has commonly relied on extracting handcrafted features from data with considerable success (Müller-Gerking et al., 1999; Van Putten et al., 2005; Ang et al., 2008). Indeed, when compared with deep learning-based methods, they have been shown to yield highly competitive performance as in Gemein et al. (2020) and Engemann et al. (2022), thus providing a valuable baseline comparison for the current study. We follow the suggestions of these authors and use a set of summary statistics which described either the time-series or power spectrum based. Specifically, these include statistical measures (standard deviation, kurtosis, skewness, mean, peak-to-peak amplitude, and quantiles), spectral measures (power in frequency bands: δ [0–2 Hz], θ [2–4 Hz], α [4–8 Hz], β_{low} [8–13 Hz], β_{mid} [13–18 Hz], β_{high} [18–24 Hz], γ [24–30 Hz], γ_{high} [30–49 Hz], and spectral entropy), and complexity measures (approximate entropy, sample entropy, SVD entropy, Hurst exponent, Hjorth parameters, line length, wavelet coefficient energy, Higuchi fractal dimension, zero crossing rate, and SVD-based Fisher information). Features were extracted using MNE. The features are extracted per-epoch and per-channel, after which they are concatenated across channels. These features are fit using a random forest with 500 estimators and a grid search over the maximum depth of a tree (4, 8, 16, None) and the maximum number of features to evaluate for splitting at each node (sqrt, log2). This method was found to work well by generating subject-level predictions by averaging the features within-subject and across epochs (Gemein et al., 2020; Engemann et al., 2022). However, as the current analyses differ in terms of dataset size and number of labels, we investigate on the training sets of both datasets whether predictive performance is higher when models are trained to predict on a per-epoch basis. In this case, the resulting predictions are averaged across epochs instead of features (as for SSL evaluation). Indeed, we find this to improve performance slightly and thus use this set-up for the final analyses.

Riemannian Filterbank. Covariance-based filterbank approaches use the spatial and spectral decomposition of EEG signals to expose the underlying neuronal activity. The application of Riemannian geometry to the covariance matrices of the EEG signals allows for an effective description of the data across different frequency bands (Sabbagh et al., 2019; 2020). Specifically, the framework corrects for distortions which arise from the linear mixing of recorded scalp-level signals arising from non-linear neural sources. The resulting features are invariant to field spread. The method assumes the covariance matrices to be full rank (Sabbagh et al., 2020), which is commonly violated, as it is here for both datasets due to the application of an average reference. In such cases of low-rank, it is suggested to project the data using principal component analysis and to only keep the components capturing the most variance. Given C EEG channels, we therefore keep $C - 1$ components. However, as we linearly interpolate bad channels for HBN and thereby reduce the rank of the covariance matrices of some subjects, we use the training set to investigate whether fewer components lead to superior performance. We do not find this to be the case and therefore use the method with $C - 1$ components for both datasets. The resulting features are fitted using a logistic regression model, using the same L2 regularization hyperparameter-grid as for SSL evaluation. Similarly to the handcrafted features, we use the training sets to compare using epoch-level or subject-level predictions and find the former to perform better on TUAB and the latter on HBN.