SELF-SUPERVISED LEARNING FOR EEG-BASED PSYCHIATRIC DISORDER CLASSIFICATION

TUUR SMOLDERS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

961677

COMMITTEE

dr. Marijn van Wingerden
Msc. Niloy Purkait

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

December 2nd, 2024

WORD COUNT

8166

ACKNOWLEDGMENTS

I would like to extend my gratitude to Dr. Marijn van Wingerden for his invaluable supervision throughout this thesis and the broader project we have worked on over the past year. The insightful feedback, patience, and support have been instrumental in enabling me to learn and, in turn, write this thesis."

SELF-SUPERVISED LEARNING FOR EEG-BASED PSYCHIATRIC DISORDER CLASSIFICATION

TUUR SMOLDERS

Abstract

Self-supervised learning (SSL) offers a solution to the scarcity of labeled electroencephalography (EEG) data, a key barrier to advancing supervised machine learning (ML) applications in psychiatric disorder classification. By pretraining encoders on unlabeled data, SSL can potentially enhance feature extraction for psychiatric disorder classification, potentially improving classification performance and advancing objective data-driven psychiatric diagnoses.

This thesis investigates four SSL implementations (within-subject and cross-subject relative positioning (RP), cross-subject shuffling (CSS), and contrastive loss) for multiclass classification of attention-deficit hyperactivity disorder (ADHD), major depressive disorder (MDD), obsessive-compulsive disorder (OCD), subjective memory complaints (SMC), and healthy controls (HCs). Traditional ML models and a deep neural network (DNN), ShallowNet, were compared, both with and without SSL-pretrained features. Results show limited efficacy of SSL: only the CSS-based RF model achieved comparable performance to the best baseline model (F1-scores: $.436 \pm .053$ vs $.446 \pm .052$, respectively).

These findings emphasize the need for selecting and fine-tuning SSL tasks aligned with specific downstream classification goals. Although current SSL implementations did not outperform baseline models, this work provides a comprehensive comparison between different SSL implementations and a foundation for future research of SSL's potential in EEG-based psychiatric disorder classification.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

All data has been acquired from the open-source Two Decades-Brainclinic Research Archive for Insights in Neurophysiology (TDBRAIN) electroencephalography (EEG) database (van Dijk et al., 2022), owned by the Brainclinics foundation, through an online request at https://brainclinics.com/

resources. The data for all participants within the dataset is anonymised. No data from human participants or animals were collected for the current thesis. Permission for the use of the dataset was granted by signing a data use agreement. All the figures in the current body of work are created and belong to the author. The thesis code can be accessed through the Github repository following the link https://github.com/TSmolders/DataScience_Thesis. Code related to the baseline models can be accessed through https://github.com/TSmolders/Internship_EEG. The code for the ShallowNet architecture has been adapted from the selfEEG Python package (Pup et al., 2024, under MIT license). Part of the preprocessing code was adapted from the pipeline provided by the authors of the TD-BRAIN dataset (van Dijk et al., 2022, under MIT license). The adapted code fragments are clearly indicated in the notebook. A generative language model (GitHub Copilot, https://github.com/features/copilot) was used for assistance with debugging and adjusting code. Code adjusted by the generative language model is clearly indicated in the notebook. In terms of writing, a generative language model (ChatGPT 4o, https://chatgpt.com/) was used to improve the author's original (!) content, for rephrasing, spell checking and grammar. The reference management software Zotero (https://www.zotero.org/) was used for storing and managing the library for the current thesis. All code was written in Python (version 3.11), see Appendix A (page 45) for the full list of utilized packages.

2 INTRODUCTION

2.1 *Problem Statement*

In current clinical practice, psychiatric disorders are diagnosed using the International Classification of Disease (ICD; Word Health Organization, 1992) and the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2013). While these frameworks offer a structured approach, they rely on subjective assessments of clinical signs and symptoms, leading to diagnostic variability and challenges in addressing overlapping conditions (Feczko et al., 2019; Fu et al., 2019; Fuchs, 2010; McTeague & Lang, 2012). The limited understanding of the etiology of psychiatric disorders further compounds the difficulty of providing precise diagnoses and effective treatments (Michelini et al., 2021). Concurrently, advancement in understanding the underlying causes of psychiatric disorders is frequently hampered by common misdiagnoses and the significant heterogeneity present in these conditions. Thus, better diagnosis of psychiatric disorders will provide insight into the pathophysiology of the disorders, addressing our lack of complete understanding of the etiology.

Electroencephalography (EEG) has emerged as a promising tool for developing objective biomarkers of psychiatric disorders, providing accessible, non-invasive insights into brain activity (Louis et al., 2016a, 2016b). In the last decade, research into EEG biomarkers in combination with supervised machine learning (ML) for a data-driven diagnostic approach has been vast (see Section 3 Literature Review). However, the reliance of on labeled EEG datasets, which are resource-intensive to create, limits the performance and generalizability of these supervised ML models.

Self-supervised learning (SSL) addresses these limitations by pretraining models on unlabeled data to extract meaningful features for downstream tasks. For instance, while labeled EEG data for ADHD patients may be limited, abundant recordings from patients with other conditions can be used for SSL pretraining. SSL can utilize all the recordings to pretrain model parameters, considering it classifies pseudo-labels independent of the disorder label. This enables models to learn transferable features, enhancing classification accuracy on labeled datasets.

While SSL has shown great promise for applications in the field of computer vision (Chen et al., 2020; Jing & Tian, 2021) and natural language processing (Devlin et al., 2019; Mikolov et al., 2013), there have been limited studies focusing on applications with EEG data. The current articles on EEG-based SSL each provide their own approach (see Section 3 Literature Review), however, to the extent of our knowledge, there currently are no papers providing a comprehensive comparison between these approaches. Furthermore, no previous literature has investigated the use of SSL for multiclass psychiatric disorder classification.

2.2 *Research Goal*

This thesis investigates the efficacy of SSL pretraining in improving EEG-based multiclass classification of psychiatric disorders, focusing on attention-deficit hyperactivity disorder (ADHD), major depressive disorder (MDD), obsessive-compulsive disorder (OCD), subjective memory complaints (SMC), and healthy controls (HCs).

Accurate classification of psychiatric disorders using SSL features can provide an objective data-driven diagnostic alternative to the current nosology, addressing the subjective assessments of clinical signs and symptoms. Concurrently, by reducing the misdiagnoses, an advancement in understanding the etiology of the psychiatric disorders can be made. In addition, the thesis will fill the current knowledge gap in the literature for the comparison of different SSL methods for multiclass psychiatric disorder classification.

We have investigated the same multiclass classification task with the same dataset in previous unpublished work, where we were able to achieve up to $.446 \pm .052$ and $.403 \pm .050$ epoch-level F1-scores with handcrafted features from closed and open eye resting-state EEG recordings respectively, using traditional machine learning algorithms with nested cross-validation. These models were only trained using the limited quantity of labeled data available (see section 4.3 Samples), providing a great opportunity for further improvement of classification using SSL pretraining with unlabeled data. Thus, these models form the state-of-the-art and baseline for the current investigation.

With the scientific & societal relevance in mind, the research questions for the current thesis are:

1. *Do SSL features pretrained using unlabeled EEG data improve classification of psychiatric disorders compared to non-pretrained ML models?*
   (a) *Which SSL pretext tasks yield the most informative features?*
   (b) *Which model architectures benefit most from SSL: traditional ML models or DNNs?*
   (c) *How well do the models classify individual disorders?*

The main findings of the thesis are that the current implementation of the SSL pretraining does not outperform the non-pretrained models. The RF trained with CSS-obtained SSL features, however, did obtain similar performance to the best baseline model (F1-scores: $.436 \pm .053$ vs $.446 \pm .052$, respectively). Comparing the predictive power of the SSL features obtained with the different pretext tasks, it is evident that it is crucial to select and tune the pretext task for the corresponding downstream multiclass classification. Moreover, the results remain consistent when comparing between the traditional ML models and ShallowNet. Notably, the class specific results show that, generally, ADHD and SMC are most accurately classified, followed by HCs. The largest distinction between the baseline models and the pretrained models are that the baseline models are additionally able to largely correctly classify MDD and OCD as well.

3 LITERATURE REVIEW

3.1 *EEG-based ML for Psychiatric Disorder Classification*

EEG-derived features have been extensively explored as biomarkers for psychiatric disorders, particularly with the rise of machine learning (ML) methods (Emre et al., 2023; Huynh et al., 2024; Mumtaz et al., 2018; Xu et al., 2019; Zhou et al., 2023). For instance, Mumtaz et al. (2018) employed

classification models such as support vector machine (SVM), logistic regression, and Naïve Bayes to distinguish between patients with MDD and HCs. These models utilized EEG-derived functional connectivity (FC) features, achieving classification accuracies as high as 98%. Additionally, Emre et al. (2023) developed multiclass classification models, including random forest (RF) and SVM, to differentiate between a range of conditions; bipolar disorder, ADHD, depression, OCD, opioid addiction, post-traumatic stress disorder (PTSD), schizophrenia, and healthy controls. Their models, trained on EEG-derived absolute power values across four frequency bands (alpha, beta, delta, theta), reached accuracies of up to 84%. Multiclass psychiatric disorder classification is particularly valuable as it aligns closely with real-world diagnostic needs, offering a holistic approach to distinguishing patients across diverse disorders using a single predictive model.

Deep learning (DL) models, such as convolutional neural networks (CNNs), offer further capabilities by automatically extracting features from raw EEG data (Lawhern et al., 2018; Zhou et al., 2023). CNN architectures, such as EEGNet (Lawhern et al., 2018), leverage convolutional filters to identify temporal and spatial patterns. By sequentially applying these filters, the CNN can capture complex brain activity features, such as the power per frequency over time. For instance, Liu et al. (2022) applied EEGNet on raw EEG data to classify MDD patients, achieving 90.98% accuracy. In turn, demonstrating the potential of DL models for extracting nuanced EEG features and/or psychiatric disorder classification.

However, despite their successes, supervised models are constrained by the need for extensive labeled datasets, which are often unavailable for EEG applications. SSL can address this limitation by pretraining models on unlabeled data to extract meaningful features for downstream tasks. While SSL has shown great promise for applications in the field of computer vision (Jing & Tian, 2021) and natural language processing (Devlin et al., 2019; Mikolov et al., 2013), there have been limited studies focusing on applications with EEG.

3.2 *Self-Supervised Learning for EEG-based Classification*

At its simplest form, SSL can be implemented using a supervised contrastive pretext task. For example, Ou et al. (2022) obtained pretrained representations of EEG data by applying a pretext task where a model learned to classify if randomly sampled segments of EEG data were in the original order or were rearranged. This pretext task was termed Temporal Rearrange (TR). The pretext task utilized an encoder that took an original or a concatenated rearranged signal as input and subsequently learned

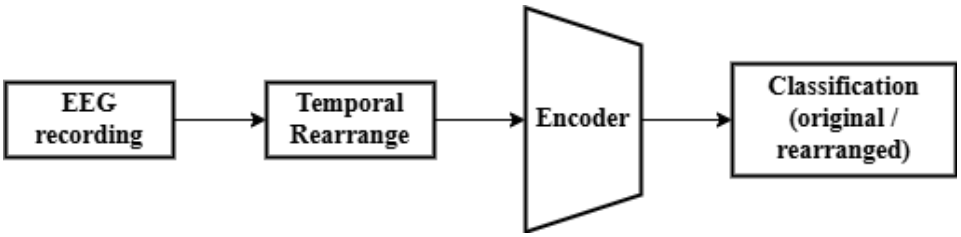

Figure 1: Diagram of pretext SSL implementation investigated by Ou et al. (2022)

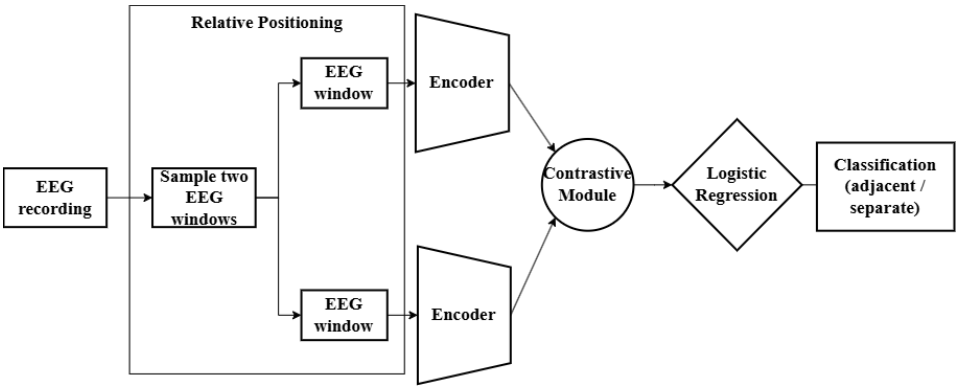

Figure 2: Diagram of pretext SSL implementation investigated by Banville et al. (2021)

to predict the correct label corresponding to the input (Figure 1). The pretrained encoder from the pretext task was then utilized for downstream Motor Imagery EEG (MI-EEG) classification. Their results show that the inclusion of SSL led to better model performance than non-pretrained models, especially when the amount of labeled data was limited (e.g. 53.00% vs 68.40% accuracy for non-pretrained model vs pretrained model using one-third of the labeled training samples).

A slightly more complex application of a supervised contrastive pretext task was investigated by Banville et al. (2021). They obtained SSL representations of EEG data by pretraining an encoder in a pretext task. One of the pretext tasks they investigated was termed relative positioning (RP). In this task, firstly, two EEG segments are randomly sampled from an EEG recording. These two EEG segments each go through the encoder separately, after which the differences in the representations of the two segments obtained by the encoder are highlighted using a contrastive module. These highlighted differences are then put through a logistic regression to classify if the two EEG segments are within or outside a given time range from each other in the original recording (Figure 2). EEG representations are obtained using the pretrained encoder from the pretext model, which are then used for two different downstream tasks; 1) sleep staging,

and 2) pathology detection. An important distinction between these two downstream tasks is that sleep staging is concerned with biological events observed within subjects (event-level), while pathology screening is concerned with patients as compared to the population (subject-level). When the amount of labeled data was reduced, Banville et al. (2021) showed that the RP pretext task, and other pretext tasks they investigated, led to better classification in downstream tasks than deep learning models that were purely trained with the labeled data or handcrafted features. For example, the SSL-learned features yielded a 22.8 point higher accuracy score as compared to a fully supervised baseline when only one example per class was available. This favor of SSL features remained up to around 10,000 examples per class. Moreover, even when all labeled data was made available, the SSL-learned feature sets were able to rival the fully supervised model and the model trained with handcrafted features. Next to improved downstream model performance, Banville et al. (2021) were able to show that the SSL features were physiologically and clinically relevant by projecting them onto a two-dimensional representation and color-coding data points for certain characteristics, such as sleep stages. Lastly, they showed that the hyperparameters of the SSL pretext task strongly influence downstream task performance. For example, the downstream pathology detection task achieved higher performance when the two EEG segments in RP were sampled from two different patients, suggesting the pretext task resulted in feature representations that were better able to distinguish between EEG recordings (subject-level). This is particularly helpful in a downstream task where the entire recording is given a single label, such as pathology detection. In turn, this shows that the chosen pretext task is crucial for downstream task performance.

Rafiei et al. (2024) provided a review of SSL implementations for EEG data. Like Banville et al. (2021), they stress the importance of selecting the right pretext task. However, they mention that considering the lack of scientific background literature, currently it is presumed a trial-and-error approach will still be required to discover the best pretext model for the downstream task. In addition, they provide an overview of different pretext data augmentations. They argue that modifying the temporal organization of the EEG in the pretext task, as done with RP, may compromise the EEG's physiological footprint, and therefore be suboptimal.

To combat this potential contamination of useful EEG information, Mohsenvand et al. (2020) surveyed neurologists to develop data augmentations that would conserve the physiological footprint of EEG recordings. Their preliminary investigation concludes with six suggested data augmentations for contrastive pretext tasks. These include, for example, a time shift where the temporal sampling in time is shifted by a predefined amount,

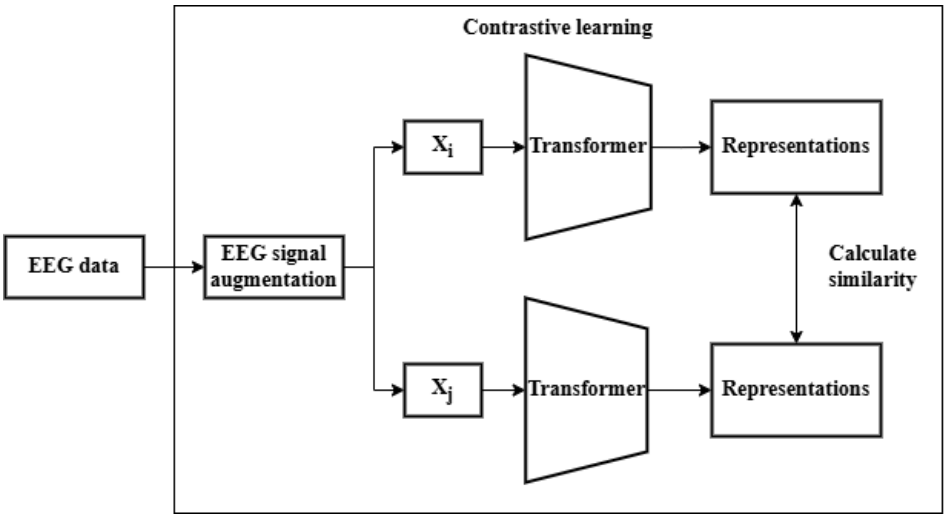

Figure 3: Diagram of pretext SSL implementation investigated by Xiao et al. (2024)

or applying a band-stop filter which significantly reduces frequencies of a specific predefined range.

There are also examples of more complex applications of SSL in context of EEG data. Xiao et al. (2024), for example, investigated self-supervised contrastive learning, with a methodology similar to Google's SimCLR (Chen et al., 2020). Here, the pretext model learns the representations of the EEG data by calculating the similarity of the representations of the EEG signals under different transformations (Figure 3). Two augmented signals (denoted as X_i and X_j in Figure 3) are put through a transformer separately, for which each a set of representations is extracted. The contrastive learning framework, subsequently, learns to minimize the distance between the representation sets extracted from two augmented signals of the same original EEG signal, and maximize the distance between representation sets extracted from two augmented signals from different original EEG signals. The weights from the transformer were then transferred to the downstream model, which was tasked with seizure detection. Xiao et al. (2024) argue that the key advantage of their contrastive learning approach is that it is able to learn more advanced features of EEG data, but at the cost of requiring a large amount of EEG data and significant computational resources. Their pretrained model did indeed outperform their non-pretrained model, for example in a within-patient seizure detection downstream task the obtained accuracy was 97.07% for the pretrained model vs 95.77% for the non-pretrained model.

3.3 *Summary*

In short, SSL provides an opportunity for obtaining informative representations using largely available unlabeled data, as opposed to being limited to training on less available labeled data. While labeled data is especially scarce for EEG data, there has been relatively little literature investigating SSL pretraining for EEG-based tasks. Nevertheless, the few articles that have investigated EEG-based SSL pretraining describe improved downstream classification performance depending on the amount of available labeled data (Table 1). Furthermore, they present a multitude of different pretext tasks (Table 1), and argue the importance of selecting the right pretext task for the corresponding downstream task, such as the difference between within-subject and cross-subject classification. Considering, the state-of-the-art model performance might be limited by the amount of available labeled EEG data (see Section **??** **??**), SSL pretraining provides an exciting opportunity for further improvement in performance of multiclass psychiatric disorder classification.

The described architecture of the pretext models and the different pretext tasks in the mentioned literature, will form the basis for the current investigation. Specifically, we will include a within-subject RP, cross-subject RP, cross-subject shuffling (CSS), and contrastive loss pretext task. The within- and cross-subject RP and the CSS pretext tasks will be based on the methodology described by Banville et al. (2021), utilizing a similar architecture with an encoder and a contrastive module. The contrastive loss pretext task will be loosely based on the methodology described by Mohsenvand et al. (2020), using the augmentations recommended by the neurologists and using a contrastive loss function. Thus, this study builds on existing work by systematically evaluating four SSL pretext tasks for multiclass classification of psychiatric disorders.

4 METHODOLOGY

4.1 *Database Description*

This study used the open-source Two Decades-Brainclinic Research Archive for Insights in Neurophysiology (TDBRAIN) EEG database (van Dijk et al., 2022), which includes raw eyes-closed (EC) and eyes-open (EO) resting-state EEG data from 1,274 participants consisting of psychiatric patients and healthy controls (620 female, age 38.67 ± 19.21, range $5 - 89$ years). The main psychiatric diagnoses include MDD ($N = 426$), ADHD ($N = 271$), SMC ($N = 119$), and OCD ($N = 75$). To facilitate correct validation and replication practices, a replication sample ($N = 106$) of the clinical and

Table 1: Overview of previous SSL literature.

Author	Contrastive method	Pretext task	n examples per class when pretrained model & non-pretrained model performance converge
Ou et al. (2022)	Supervised learning	Temporal Rearrange	NA*
Banville et al. (2021)	Feature contrast	Relative Positioning,	10^3
		Temporal Shuffling,	10^3
		Contrastive Predictive Coding.	10^3
Mohsenvand et al. (2020)	Contrastive loss	Neurologist recommended augmentations	NA**
Xiao et al. (2024)	Contrastive loss	Gaussian noise + Cut and rearrange	Not reached

**The article does not include an ablation analysis for more than 50% of labeled samples, however, at 50% labeled data converge of performance is not close to being reached*

***The article does not include an ablation analysis, however, at 100% of labeled samples model performance of pretrained and non-pretrained models are the same.*

treatment outcome data was blinded by the authors of the TDBRAIN database. Unfortunately, the replication sample is not suitable for out-of-sample evaluation of our current multiclass classification task, and was therefore not used to evaluate the current models. However, we were able to utilize the EEG data of the replication sample during the pretext SSL task. Further detail on what parts of the data were used for the current analysis will be described in Section 4.3 Samples.

The EEG data was collected during one or multiple sessions of a two-minute resting-state recordings with EC or EO. EEG recordings included 26 electrodes following the 10-10 international system and auxiliary electrodes for electromyography (EMG), electrocardiography (ECG), and eye movements. Recordings were sampled at 500 Hz, with a 100 Hz low-pass filter applied before digitization. For further experimental conditions refer to van Dijk et al. (2022).

An additional spreadsheet is provided containing the clinical, demographic, and practical information for all participants. The recorded diagnoses within the TDBRAIN database are distinguished as either a formal diagnosis, confirmed by a licensed clinician, or as a referral-indication, denoting the participant was referred by a general practitioner or psychologist/psychiatrist based on an indication of that particular disorder. In total, 39 single or comorbid psychiatric diagnoses, as well as unknown/missing entries and HCs are included. Healthy participants were denoted with the value 'HEALTHY' and unknown or missing entries with the value 'UNKNOWN' or NaN. Participants diagnosed with a psychiatric disorder were denoted with their respective diagnosis as the value, for example, participants with MDD were denoted with 'MDD'. In the case of comorbidity, all respective disorders were included in the value (e.g. 'MDD/ADHD')

All EC resting-state EEG data was used for the pretext task ($N = 1,274$), while only the EC resting-state EEG data labeled with the selected psychiatric disorders for classification was used for the downstream task ($N = 938$). For all participants, only the recording from their first session was used. Class imbalance in the labeled sample was dealt with by subsampling. No separable features, such as age or gender, were used.

4.2 *Data Analysis Pipeline*

A schematic overview of the analysis pipeline can be seen in Figure 4. The steps within this pipeline will be discussed in further detail in the following sections.

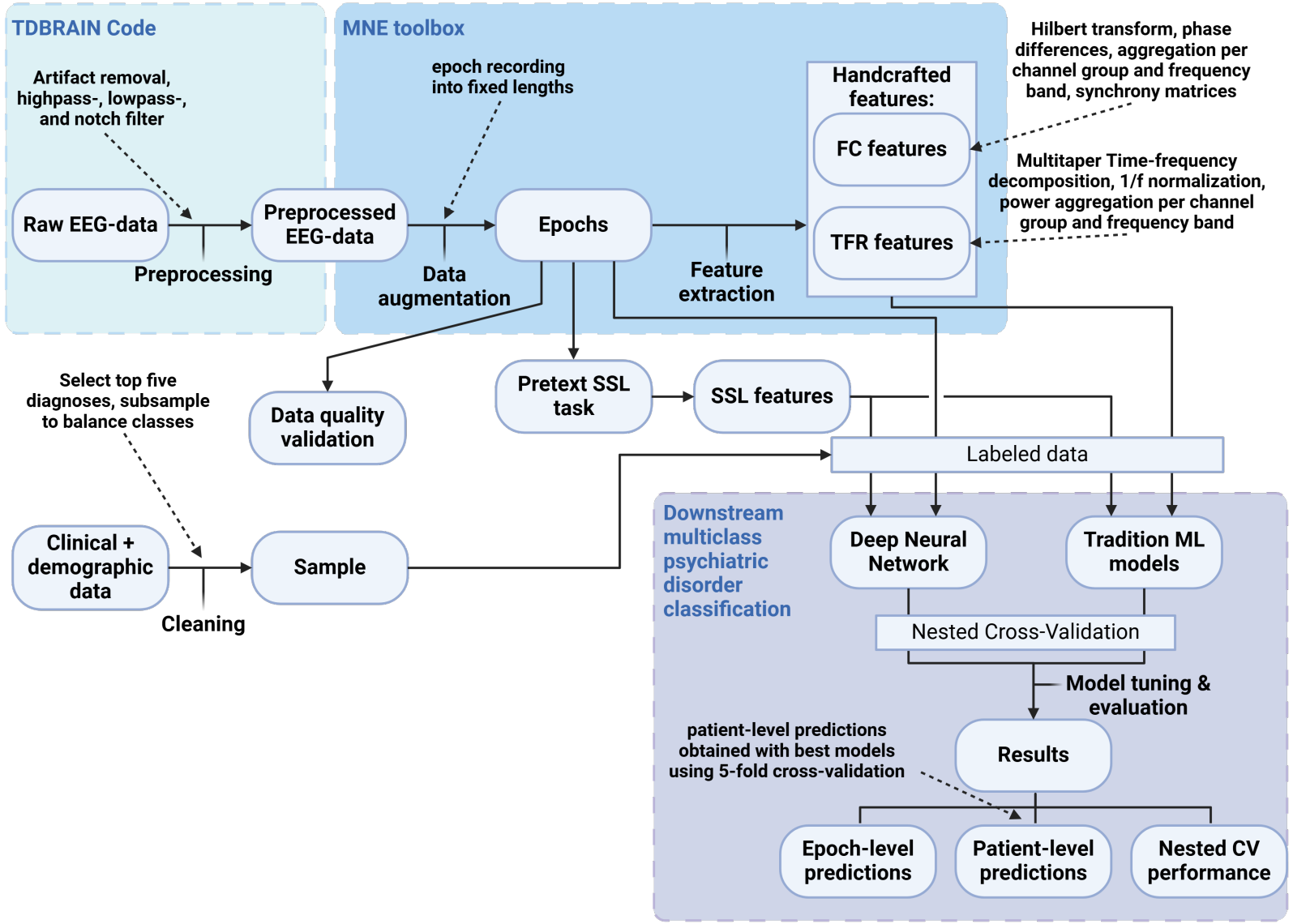

Figure 4: Schematic overview analysis pipeline. EEG = electroencephalography; TFR = time-frequency representation; FC = functional connectivity; ML = machine learning; SSL = self-supervised learning; cv = cross-validation

4.3 *Samples*

Two separate samples were created for the SSL pretraining and the downstream classification. The SSL pretraining subset consisted of all eyes-closed (EC) EEG recordings from the first session of 1,274 participants, regardless of diagnosis. To maiximize the sample size for the downstream classification subset, an aggregated 'diagnosis' variable was created by using the formal diagnosis where available and the referral indication otherwise. The downstream subset was subsequently restricted to five target classes: MDD ($N = 323$), ADHD ($N = 176$), SMC ($N = 119$), OCD ($N = 49$), HC ($N = 47$). Participants were evenly sampled across classes to minimize the risk of prediction bias in the models, resulting in 45 participants per class. Balancing was conducted after handcrafted feature extraction, to ensure that the subsampled entries had no to little missing variables. Due to the exclusion of two unusable artifact-ridden EEG recordings, the final downstream sample consisted of 45 participants per class (Table 2)

Table 2: Sample Characteristics by Diagnosis

Diagnosis	*N*	Age (mean ± SD)	Gender (% male)	Formally diagnosed (%)
ADHD	45	23.7 ± 14.2	64.4	37.8
HC	45	32.3 ± 13.9	35.6	100
MDD	45	42.7 ± 13.5	37.8	46.7
OCD	45	32.5 ± 11.5	60.0	82.2
SMC	45	63.5 ± 7.3	37.8	0

Note. ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

4.4 *Preprocessing*

The raw EEG data was preprocessed using the pipeline provided by the authors of the TDBRAIN dataset (van Dijk et al., 2022). Firstly, according to the method of Gratton et al. (1983), eye-blink artifacts were eliminated from the EEG signal by computing and removing the bipolar electrooculography (EOG). The data was, subsequently, bandpass-filtered between 0.5 to 100 Hz, and notch-filtered at 50 Hz. Next, the following artifacts were detected in the filtered signal: EMG, sharp channel-jumps, electrode bridging, extreme correlations, extreme voltage swings, kurtosis, and residual eyeblinks. If more than two-thirds of the signal from an EEG electrode

contained detected artifacts, the electrode's signal was repaired using a Euclidean distance weighted average of at least three neighboring electrodes. If repair was not possible, the entire recording was excluded.

4.5 *Data Augmentation*

The preprocessed EEG recordings, each lasting approximately 120 seconds, were divided into twelve 9.95-second epochs. This segmentation increased sample size and minimized signal non-stationarity. To ensure reliability in wavelet and multitaper analyses, it is crucial that the signals are stationary (Cohen, 2019). Due to the nature of the resting-state recordings, the data could be divided into epochs of a fixed length without splitting task-evoked brain activity. The 9.95-second length was chosen for two reasons: First, it allows for at least three full cycles of a 1 Hz frequency, necessary for accurate time-frequency analysis. Second, it consistently yields twelve epochs from each recording, providing uniformity across the dataset.

4.6 *Data Quality Validation*

Prior to the investigation, a RF classifier was trained on Power Spectral Density (PSD) features derived from the epoched EEG data to classify the epochs as either EO or EC. Previous research has demonstrated that EEG power levels differ significantly between EO and EC conditions, particularly in the alpha band (Barry et al., 2007; Li et al., 2009). Therefore, assessing the classifier's ability to accurately distinguish between EO and EC conditions provided a means to validate the quality of the preprocessed data. The average 5-fold cross-validation (CV) accuracy and feature importances were evaluated to ensure the data quality and the expected relevance of alpha-band power. The classifier achieved an average accuracy 74.7% (mean accuracy = $.747 \pm .016$) in distinguishing EO and EC epochs. Additionally, alpha-band PSD features consistently emerged as the most important across all CV folds, confirming that the preprocessed data contained valuable information on brain activity and was of high quality.

4.7 *Handcrafted Features*

The non-pretrained traditional ML models (SVM, RF, GBC) were trained with two types of handcrafted features: 1) statistical features (mean, standard deviation, median, skewness, kurtosis) from a time-frequency representation (TFR) of the epoch, and 2) functional connectivity (FC) features of the synchrony between electrode groups. All features were aggregated over

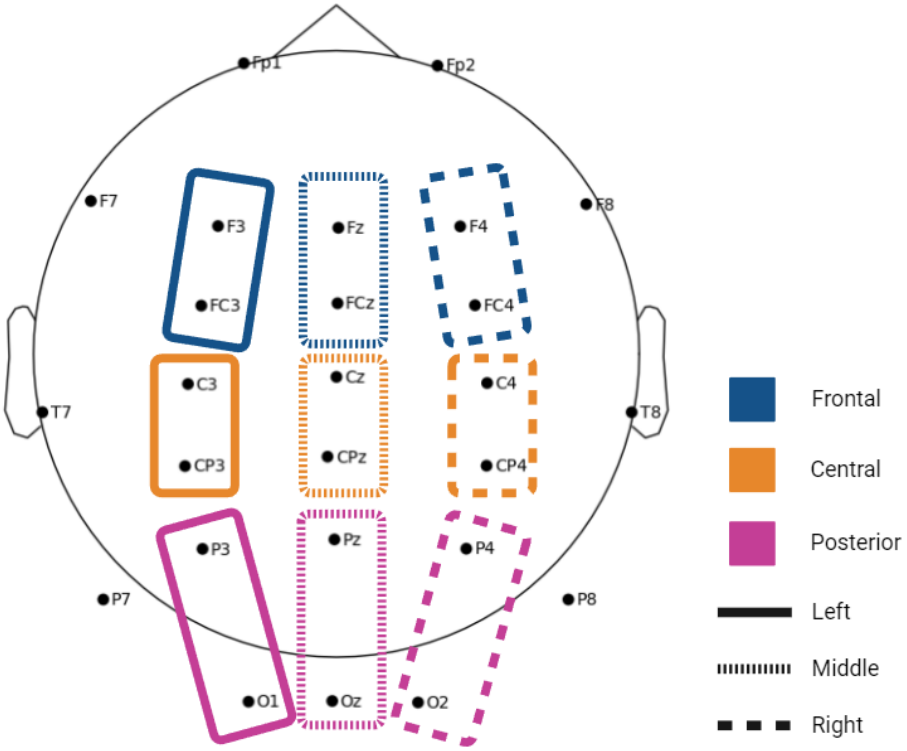

Figure 5: Electrode montage and groups. The figure depicts a top-down view of a schematic head, oriented with the face pointing north. Electrodes included in the analysis are grouped and highlighted with distinct borders, representing their respective electrode groups. These groups are labeled based on their location on the head: Frontal, Central, or Posterior, combined with their lateral positioning as Left, Middle, or Right, forming nine distinct electrode groups. Electrodes that do not belong to these nine groups were excluded from further analysis.

Table 3: Frequency Bands and Frequencies of Interest

Frequency band	Lower value – upper value (Hz)	Selected frequencies (Hz)
Delta	1 – 4	1, 1.5, 2, 2.5, 3
Theta	4 – 8	4, 4.75, 5.5, 6.25, 7
Alpha	8 – 13	8, 9, 10, 11, 12
Beta	13 – 30	13, 17.25, 21.5, 25.75, 30
Gamma	30 – 90	42, 54, 66, 78, 90

Note. The lower and upper values of each frequency band are given, as well as the frequencies selected within each frequency band for the time-frequency representation.

electrode groups (Figure 5) and frequency bands (Table 3), with electrodes outside predefined groups excluded. The next subsections will detail the feature extraction methods for the two feature types.

4.7.1 *Statistical Time-Frequency Representation Features*

The TFR for each epoch was computed using the multitaper method, which is effective for non-phase-locked, non-time-locked data and helps smooth noise while preserving key features (Babadi & Brown, 2014). Smoothing noise and highlighting broader signal features are especially beneficial for analyzing resting-state EEG data, which often lacks distinct events. The multitaper method used multiple discrete prolate Slepian sequences (DPSS), which are orthogonal tapers, to taper the signal in various ways (Babadi & Brown, 2014). A fast Fourier transform (FFT) was applied to the tapered signals to produce a power spectrum, which was then averaged over frequency-specific temporal windows. The resulting TFR was normalized to a relative TFR by dividing each sample by the sum of the average power over time per frequency in the epoch, ensuring the average power across all frequencies and time points summed to 1. The TFR was aggregated over electrode groups (Figure 5) and frequency bands (Table 3), and statistical features (mean, standard deviation, median, skewness, and kurtosis) were calculated for each electrode group and frequency band. This process resulted in 225 statistical TFR features per epoch.

4.7.2 *Functional Connectivity Features*

The consistency of the phase relation between two signals can provide an estimate of synchrony, which in turn reflects FC. In this study, synchrony was estimated between each electrode group (Figure 5) and across frequency bands (Table 3), resulting in a total of 405 FC features. The

signals were band-pass filtered according to the frequency range of each band, and the synchrony estimates were determined by the difference in phase of two signals, computed with the Hilbert transformation (Hilbert, 1912). Synchrony estimates were averaged within each electrode group (Figure 5). Lastly, only one direction of bidirectional synchrony was used, and self-connections were excluded, resulting in 180 FC features per epoch.

4.8 *Feature Selection*

Feature extraction resulted in a high-dimensional feature set of statistical TFR and FC features, totaling 405 features. To mitigate the risk of overfitting, Boruta feature selection was applied before training the ML models. Boruta is an "all-relevant" feature selection method, which identifies all features that contribute meaningfully to the model, unlike "minimal-optimal" methods that focus only on the most predictive features (Kursa & Rudnicki, 2010). This approach is particularly useful when the goal is to explore mechanisms related to the research topic. The 'maximum depth' and the 'number of estimators' hyperparameters for the Boruta-wrapped RF were tuned by identifying the values at which a RF classifier achieved convergence with the lowest Out-of-Bag (OOB) error, as outlined by Kursa and Rudnicki (2010). This process resulted in the selection of 319 relevant features from the original 405. For a comprehensive list of the selected features, refer to Appendix B (page 46).

4.9 *Self-Supervised Learning Features*

SSL features were learned using four pretext tasks (see Section 4.10 Pretext Tasks). Each of these pretext tasks included the same ShallowNet architecture as a learnable encoder (Figure 6). ShallowNet contained a temporal convolutional layer followed by a spatial convolutional layer, applying filters across the electrodes. The entire architecture is shown in Figure 6. The SSL features were outputted by the final linear dense layer of 100 neurons, and subsequently used for the pretext and downstream tasks. By pretraining the encoder, ShallowNet was designed to learn informative features for multiclass psychiatric disorder classification.

After pretraining, the ShallowNet encoder was used to extract the SSL features from the labeled data, and subsequently used as the input to the downstream traditional ML models. An exception was made for the downstream ShallowNet model, where instead the layers of the pretrained ShallowNet were frozen and a trainable multiclass classification head was added.

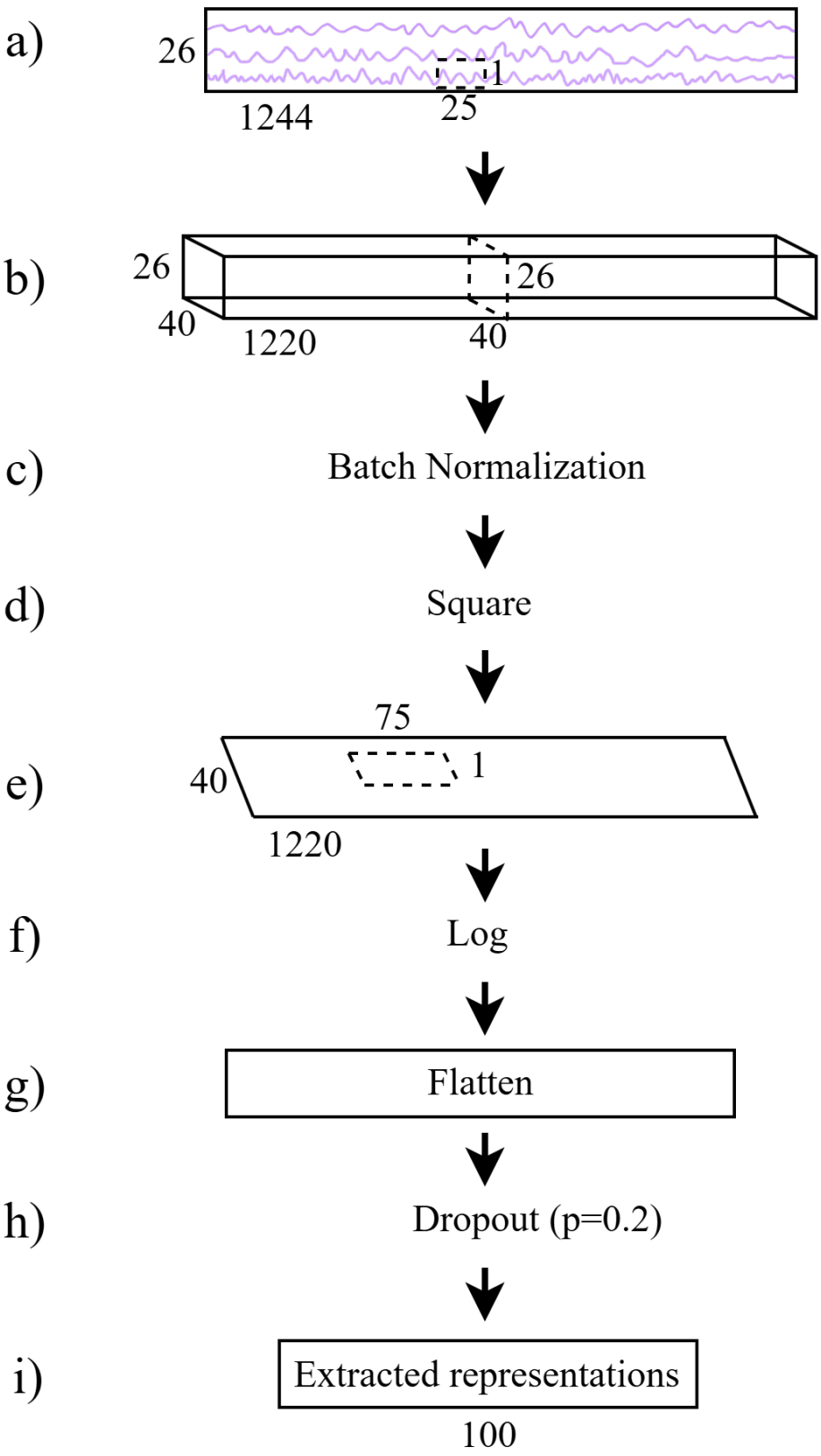

Figure 6: Diagram of the ShallowNet architecture. Starting with the raw EEG data, a temporal convolution $(1, 25)$ is applied (a), followed by a spatial convolution $(26, 1)$ across the electrodes (b). The convolutional layers are followed by a batch normalization (c) and a squaring non-linearity (d), average pooling $(1, 75)$ (e), a logarithm non-linearity (f), flattening layer (g), dropout($p = 0.2$) (h), and finally a linear output layer outputting 100 representations (i).

ShallowNet was chosen over other CNN architectures, due to its lower computational demands and fewer parameters. EEGNet (Lawhern et al., 2018), another shallow CNN with relatively little model parameters, was considered in preliminary tests, early results (not shared) indicated that ShallowNet performed slightly better as the pretrained encoder for the downstream task.

4.10 *Pretext Tasks*

Four pretext tasks were used to explore different approaches for extracting SSL features: within-subject relative positioning (RP), cross-subject RP, cross-subject shuffling (CSS), and contrastive loss with neurologist-recommended augmentations. Table 4 provides an overview of the main parameters for each task, including sample size, training epochs, batch size, and pretext model performance. The number of training epochs and batch sizes for the pretext models were determined through preliminary testing (results not shared), where hyperparameters were manually tweaked based on the downstream performance of a simple SVM without extensive optimization. The goal was to identify working hyperparameters rather than optimal ones. Each pretext task is described in more detail in the following subsections.

4.10.1 *Relative Positioning*

Similar to the pretext task described by Banville et al. (2021), during the RP pretext task, the model learned to classify if two randomly sampled epochs were adjacent or separate from each other based on their index in their original recording (Figure 7). The pretext model was able to learn this by obtaining 100 representations for each sampled epoch with a learnable encoder; ShallowNet (Figure 6). Using a contrastive module, the contrast between the two representations sets are computed, which is finally put through a single dense layer for binary classification (Figure 8. The predicted labels are compared to the actual labels using a logistic loss:

$$\ell_{\log} = \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-y_i \hat{y}_i)) \tag{1}$$

Where $\mathcal{Y} = \{-1, 1\}$ is the true label, $\hat{y}_i$ the predicted label for the i-th sample, and N the total number of samples. The logistic loss was implemented considering the proven effectiveness in the investigation of Banville et al. (2021).

Epoch pairs were pseudo-labeled as "adjacent" or "separate" based on their indices within the recording. The T_{pos} hyperparameter determined

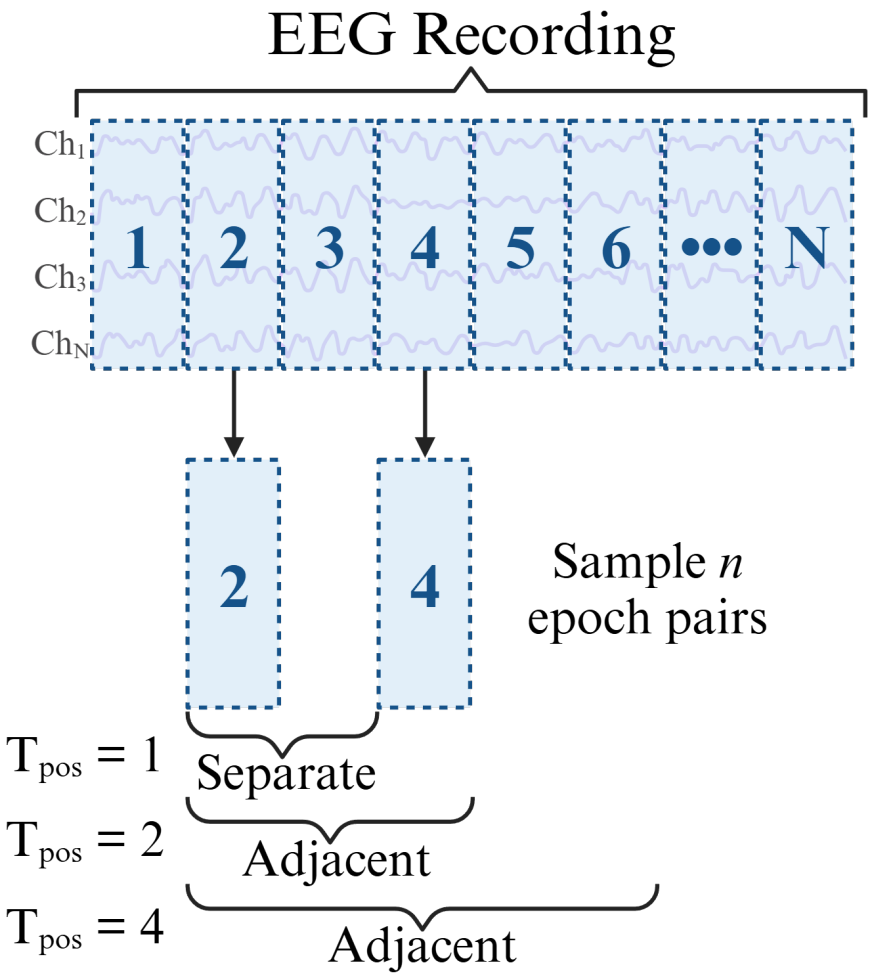

Figure 7: Visual explanation of the relative positioning pretext task. T_{pos} determined the maximum difference between the indices of two epochs for which the epoch pairs would be pseudo-labeled as adjacent

Table 4: Parameters for pretext RP tasks

Pretext task		n EEG epoch pairs	n model training epochs	batch size	F1-score
Within-Subject Relative Positioning	T_{pos} 1	27,380	300	250	.523
	T_{pos} 2	52,272	300	250	.516
	T_{pos} 4	69,720	300	250	.529
Cross-Subject Relative Positioning	T_{pos} 1	1,293,906	30	1,000	.399
	T_{pos} 2	1,293,906	30	1,000	.399
	T_{pos} 4	1,293,906	30	1,000	.459
Cross-Subject Shuffling		150,156	50	1,000	.890
Contrastive Loss		1,244	3,000	1,244	NA

Note. The sample size (n EEG epoch pairs), the training epochs, the batch size, and the F1-score on the test set per specific pretext task is given. The F1-score is obtained from the predictions of the pretext models on the test set corresponding to the pretext task. No F1-score was obtained for the contrastive loss pretext task, considering the architecture does not predict any labels.

the maximum index difference for which pairs were considered adjacent (Figure 7). This T_{pos} hyperparameter was varied for a value of one, two, and four. With $T_{pos} = 1$, for example, epochs were adjacent only if their indices were immediately next to each other. With $T_{pos} = 4$, epochs could be adjacent even if their indices differed by up to four positions. It was expected that the RP task would be more difficult with a higher T_{pos} value, considering the physiological difference between the adjacent and separate pairs becomes smaller when the range between the pairs are increased.

Two RP pretext tasks were used: within-subject RP and cross-subject RP. The model's architecture and pseudo-labeling were the same for both tasks, but within-subject RP sampled pairs from the same participant, while cross-subject RP sampled pairs from different participants. The latter was expected to be more challenging due to the physiological differences across subjects, which might overshadow the temporal differences between epochs. Nevertheless, Banville et al. (2021) showed that the cross-subject RP pretext task was particularly successful for downstream patient-level classification, such as pathology-detection.

For the within-subject RP task, given T_{pos}, related to the positive context around each epoch x_i, we sample N labeled pairs per recording (cf. Banville et al. 2021):

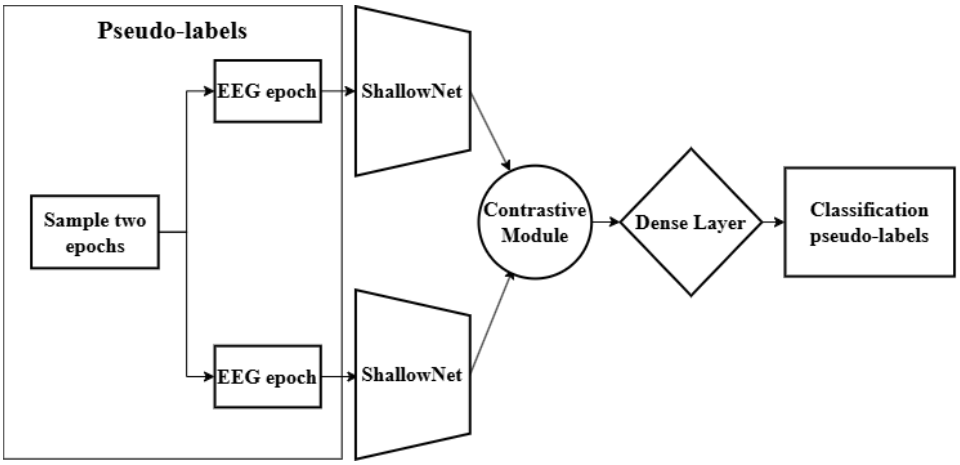

Figure 8: Schematic overview contrastive module pretext tasks. This diagram is a schematic overview of the architecture of the pretext model used for the two different relative positioning and cross-subject shuffling pretext tasks. The sole difference between these three pretext tasks are the created pseudo-labels for classification

$$\mathcal{Z}_N = \left\{ \left((x_{t_i}, x_{t'_i}), y_i \right) \;\middle|\; i \in [N],\ (t_i, t'_i) \in \mathcal{T},\ y_i \in \mathcal{Y} \right\}, \tag{2}$$

where $\mathcal{Y} = \{-1, 1\}$ and $\mathcal{T} = \{(t, t') \in \frac{\lfloor \frac{M}{T} \rfloor !}{2!(\lfloor \frac{M}{T} \rfloor - 2)!}\}$. $\mathcal{T}$ is the set of all pairs of epoch indices (t, t') obtainable from epochs of size T in a recording of size M without repetition. $y_i \in \mathcal{Y}$ is determined by the T_{pos} parameter:

$$y_i = \begin{cases} 1, & \text{if } |t_i - t'_i| \leq \tau_{\text{pos}} \\ -1, & \text{if } |t_i - t'_i| > \tau_{\text{pos}} \end{cases} \tag{3}$$

For the cross-subject RP task, $y_i \in \mathcal{Y}$ is determined the same. However, considering the pairs are sampled across participants, we sample N labeled pairs per two recordings. Given equation 2, here $\mathcal{T} = \{(t, t') \mid t \in [N_1], t' \in [N_2], t < t'\}$, where $N_i = \left\lfloor \frac{M_i}{T} \right\rfloor$ is the number of epochs per recording. Thus, here $\mathcal{T}$ is the set of all pairs of epoch indices (t, t') obtainable from epochs of size T between two recording of sizes M_1 and M_2 without repetition and self-pairs (e.g. $[t = 2, t' = 2]$).

To balance the class distribution, the number of 'separate' pairs was subsampled to match the 'adjacent' pairs, particularly for lower T_{pos} values. Due to computational constraints, the number of cross-subject epoch pairs was reduced by two pairs per label for each participant combination. For sample sizes, see Table 4.

4.10.2 *Cross-Subject Shuffling*

Cross-subject shuffling utilized the same pretext model architecture as the RP pretext tasks (Figure 7. The main difference between CSS and RP is the pseudo-labeling. In CSS, the pairs of epochs are either sampled from the same participant or from different participants, which the pretext models learns to classify. Thus, given equation 2, the set of possible epochs pairs is defined as:

$$\mathcal{T} = \bigcup_{r \in \mathcal{R}} \{(t, t') \mid t, t' \in [N_r],\ t < t'\} \cup \bigcup_{\substack{r, r' \in \mathcal{R} \\ r \neq r'}} \{(t, t') \mid t \in [N_r],\ t' \in [N_{r'}],\ t < t'\}. \tag{4}$$

Where $\mathcal{R}$ is the set of all recordings, $N_r = \left\lfloor \frac{M_r}{T} \right\rfloor$ is the number of epochs in recording r, with M_r being the size of recording r and T being the epoch size. $\bigcup_{r \in \mathcal{R}} \{(t, t') \mid t, t' \in [N_r],\ t < t'\}$ samples all pairs of epoch indices (t, t') within the same recording r, ensuring $t < t'$ to avoid duplicate pairs (e.g. $[t = 7, t' = 8]$& $[t = 8, t' = 7]$). $\bigcup_{\substack{r, r' \in \mathcal{R} \\ r \neq r'}} \{(t, t') \mid t \in [N_r],\ t' \in [N_{r'}],\ t < t'\}$ samples all pairs of epoch indices (t, t') between different recordings r and r', constrained by $r \neq r'$ and $t < t'$ to avoid self-pairs and duplicate pairs. To balance the sampled labeled epoch pairs, the number of cross-subject pairs were limited to the maximum amount of possible epoch pairs within a recording (66), resulting in a total of 150,156 epoch pairs (Table 4).

The CSS task was chosen due to the potential advantages shown by cross-subject pretext tasks for downstream patient-level classification (Banville et al., 2021). Unlike the cross-subject RP task, the positions of the sampled epochs within the recordings were not important in CSS, which was expected to make the task easier to learn. Moreover, since the position of the epochs is irrelevant for downstream psychiatric disorder classification, the CSS task was anticipated to extract more relevant features for the downstream task compared to the cross-subject RP pretext task.

4.10.3 *Contrastive Loss*

The architecture of the model in the contrastive loss pretext task differs from the RP and CSS tasks (Figure 9. Here, pairs of randomly augmented versions are created from each epoch. The two different augmented versions sharing the same original epoch all form positive pairs, while pairs of augmented versions that do not share the same original epoch all form negative pairs. Negative pairs are not explicitly sampled. Instead, given a positive pair, similar to Chen et al. (2020), all other $2(N-1)$ augmented

examples within a batch define the negative pairs. Representations from each augmented epoch are extracted using a learnable ShallowNet encoder, then passed through a multilayer perceptron (MLP) projection head with ReLU activation. The contrastive loss is computed using the *NT-Xent* (Normalized Temperature-Scaled Cross Entropy) loss function. As described by Chen et al. (2020), the loss function for a positive pair of examples (i, j) is defined as:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{\{k \neq i\}} \exp(\text{sim}(z_i, z_k)/\tau)}, \tag{5}$$

Where $\text{sim}(\mathbf{u}, \mathbf{v})$ is the cosine similarity between the representations, and τ is a temperature parameter that controls the influence of similar versus dissimilar pairs. A temperature value of 0.1 was used, based on previous studies (Chen et al., 2020; Mohsenvand et al., 2020).

This loss function is calculated across all positive pairs in a batch, both (i, j) and (j, i). Intuitively, the contrastive loss calculates the distance between the representations of positive and negative pairs, and penalizes the model to learn to minimize the distance in the representations of positive pairs, and maximize the distance in the representations of negative pairs.

For each batch, $1,244$ positive pairs were created from all epochs, and $2,486$ negative pairs were formed by considering all other augmented examples in the batch. This pretext task was chosen based on its successful application in both computer vision (Chen et al., 2020; Jing & Tian, 2021) and EEG-based tasks (Mohsenvand et al., 2020; Xiao et al., 2024), despite its higher computational demands.

Our approach of the contrastive loss pretext task, combines the described methodologies of Chen et al. (2020) and Mohsenvand et al. (2020). Augmentations applied to the original epochs were inspired by Mohsenvand et al. (2020), who consulted EEG experts to identify augmentations that preserved the interpretability of the data. Four augmentations were selected for this task. Similar to Chen et al. (2020), the augmentations were applied sequentially with randomly sampled values within predefined ranges (Table 5).

4.11 *Cluster Analysis*

Cluster analyses were conducted to explore potential class-specific clusters within the handcrafted and SSL feature sets. Clear separation between clusters would indicate distinct feature patterns for each psychiatric disorder, suggesting that the models could effectively differentiate between the classes.

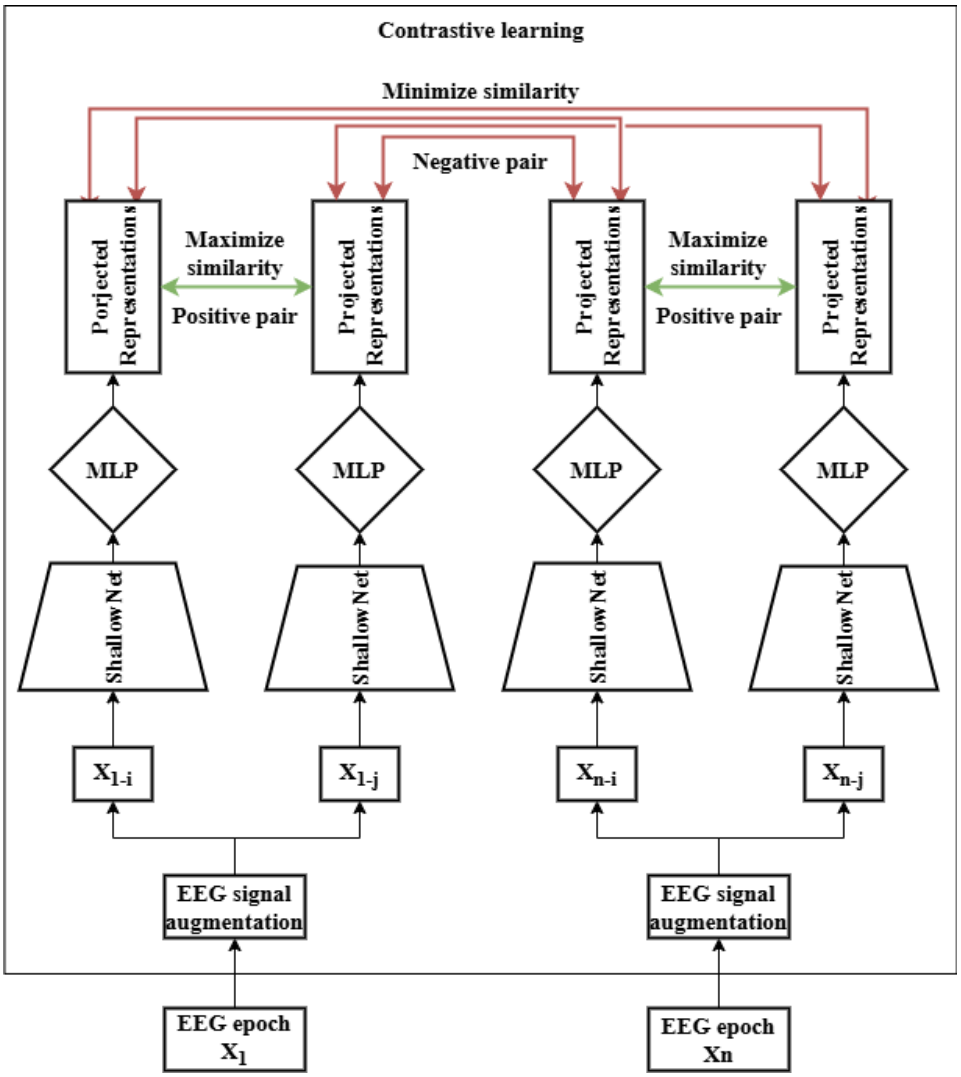

Figure 9: Schematic overview of contrastive loss pretext task. For n epochs (X_n), two random augmented versions are created $(X_n - i, X_n - j)$. A ShallowNet encoder extracts representations from all augmented versions, which are projected using a multilayer perceptron (MLP). Using a contrastive loss, the model subsequently learns to minimize the similarity between the representations of augmented epochs sharing the same original epoch (positive pair), and maximize the similarity between the representations of augmented epochs with different original epochs (negative pair).

Table 5: Augmentations for contrastive loss

Augmentation	Lower value – upper value
Amplitude scale	0.5 – 2
Additive Gaussian noise	5 – 10
Band-stop filter	4.8 – 59
Zero-masking	0 – 150

Note. These four augmentations are applied, in order of the table, to the EEG epochs. Each EEG epoch is augmented twice with randomly selected values within the given ranges. For amplitude scale, the amplitude of the original epoch is multiplied by the selected value. Afterwards, a random amount of Gaussian noise is added by adding a matrix of the same shape as the epoch containing random values from a uniform distribution multiplied by the randomly selected value. Next, a band-stop filter is applied, removing the frequencies with a range of two on either side of the randomly selected value. Lastly, the randomly selected value for the zero-masking determines the length of a randomly selected place in the epoch to be zero-masked.

t-distributed Stochastic Neighbor Embedding (t-SNE), a non-linear dimensionality reduction method, was used for the analysis. t-SNE is known for its superior visualization capabilities compared to other methods, such as Stochastic Neighbor Embedding, due to its reduced tendency to crowd points in the center of the map (Cieslak et al., 2020; Maaten & Hinton, 2008; Taskesen & Reinders, 2016).

Both 2D and 3D t-SNE visualizations were generated for various perplexity values. Perplexity acts as a smooth estimate of the effective number of neighbors (Maaten & Hinton, 2008). The visualizations were color-coded by class to assess clustering patterns and their potential pathophysiological relevance. However, none of the visualizations revealed clear class-specific clustering (Appendix C, page 46), suggesting challenges for the models in distinguishing between classes based on the features provided.

4.12 *Downstream Model Development*

The downstream multiclass classification task utilized three traditional ML models (SVM, RF, GBC) and ShallowNet. The architecture of the downstream ShallowNet model is described in Section 4.9 Self-Supervised Learning Features, with the key difference being the addition of a classification head. This head includes a batch normalization layer, dropout,

and dense layers with ReLU activation. The number of dense and dropout layers was optimized during hyperparameter tuning. For the downstream ShallowNet, raw EEG data was used as input, rather than the extracted feature sets. Unlike the pretrained ShallowNet, the baseline ShallowNet was not pretrained in any SSL pretext task. For pretrained models, the layers from the pretext ShallowNet were transferred and frozen, leaving only the classification head trainable.

The traditional ML models employed in the current study were a SVM, a RF, and a GBC. These models were chosen to encompass distinct algorithmic approaches to prediction, specifically margin-based learning and ensemble learning. This selection allowed for a comprehensive investigation of the strengths and limitations of different ML techniques within the analysis. The choice of these algorithms was additionally based on their demonstrated effectiveness in previously reported EEG-based prediction problems (Mari et al., 2022; Saeidi et al., 2021). Each ML model was trained separately with the Boruta-selected handcrafted features and each different SSL feature set, pretrained with the different pretext tasks.

Before training, features were standardized by subtracting the mean and dividing by the standard deviation, calculated only from the training data to prevent test set leakage. Missing feature values from twelve epochs, due to a 'bad' EC recording, were imputed with the mean for each feature.

4.13 *Hyperparameter Tuning & Model Evaluation*

Hyperparameters for each downstream model were optimized using nested cross-validation (CV) to minimize bias. The nested CV consisted of three inner splits and five outer splits. Stratified and grouped K-fold splitting was applied to ensure balanced class distribution and that all epochs from a participant remained in the same split. In the inner loop, hyperparameters were tuned by maximizing cross-validated F1-scores. A grid search was used for SVM, and a randomized search was applied to RF, GBC, and ShallowNet models due to computational constraints. Hyperparameter ranges are shown in Table 6.

Table 6: The hyperparameters of the downstream models and the search ranges during optimization.

Algorithm	**Hyperparameter search ranges**		
SVM	*C*: [0.01, 0.1, 1, 10, 100]	*gamma*: [0.1, 1, 10, 100]	*kernel*: ['rbf']
RF	*n estimators*: [200, 400, 600 ... 2000]	*max. features*: ['log2', 'sqrt']	*max. depth*: [10, 20, 30 ... 110, None]
	min. samples per split: [2, 5, 10]	*min. samples per leaf*: [1, 2, 4]	*bootstrap*: ['True', 'False']
GBC	*n estimators*: [100, 200, 300 ... 1000]	*loss*: ['log_loss']	*learning rate*: loguniform(1e-3, 1e-1, 1000)*
	subsample: [0.5, 0.6, 0.7, ... 1]	*criterion*: ['friedman_mse']	*min. samples per split*: [2, 5, 10]
ShallowNet	*learning rate*: loguniform(1e-5, 1e-1, 1000)**	*optimizer*: ['Adam', 'RMSprop', 'SGD']	*dense layers*: [1, 2, ... 5]
	batch size: [50, 100, 178, 350, 476, 700]	*dropout chance*: [0 ... 0.8]	

Note. The hyperparameters of the SVM were tuned using a grid-search approach, while the hyperparameters of the RF, GBC, and ShallowNet were tuned using a randomized-search approach. SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier. *The search range for the learning rate hyperparameter of the GBC is a loguniform distribution of 1000 bins within 1e-3 to 1e-1. **The search range for the learning rate hyperparameter of the ShallowNet is a loguniform distribution within 1e-5 to 1e-1.

For pretext models, no hyperparameter optimization or nested CV was performed. The TDBRAIN replication sample was used as the test set, while the remaining data was split into training and validation sets with the same stratified and grouped splitting strategy as the downstream models. The training set was used for model fitting, and the validation set was used to assess generalization by comparing F1-scores and loss during training.

Downstream models were evaluated using the mean F1-score from the five outer folds. A one-tailed paired t-test was conducted to compare the F1-score distribution between non-pretrained and pretrained models, with the alternative hypothesis that the pretrained models would perform better.

For each model, predictions on the entire dataset were obtained with the tuned model with the highest F1-score within the nested CV to create a confusion matrix and to calculate the precision, recall and F1-score for each class. A non-nested 5-fold CV was employed to obtain the predictions, ensuring that each sample was part of exactly one test set, with predictions based on the model retrained on the corresponding training set. The same five outer splits used in the nested CV were applied. This process aimed to assess the model's ability to accurately predict specific psychiatric disorders based on the underlying class predictions.

4.14 *Patient-Level Predictions*

For the best traditional ML model trained with handcrafted features, the non-pretrained ShallowNet, and the best model trained with SSL features, patient-level predictions were obtained by summing the class probabilities across epochs for each participant. Epoch-level predictions were derived using the non-nested 5-fold CV approach, with class probabilities computed for each epoch rather than direct class predictions. The class with the highest summed probability across the epochs of each participant was used to determine the patient-level prediction.

Patient-level predictions were then used to generate a confusion matrix and calculate precision, recall, and F1-score for each class, assessing the model's ability to accurately predict psychiatric disorders at the patient level.

5 RESULTS

In the following section, the results regarding the investigation of SSL features for multiclass psychiatric disorder classification will be described. First, an overview of the epoch-level performance of all models will be investigated, after which the patient-level performance, in addition to the class-specific results, of a few selected models will be described in more detail.

5.1 *Epoch-level model performance*

Table 7: Epoch-level model performance in F1-score

Condition		SVM	RF	GBC	ShallowNet
Non-pretrained		.091 ± .019	.410 ± .033	**.446 ± .052**	**.451 ± .084**
Within-Subject Relative Positioning	T_{pos} 1	.228 ± .033**	.325 ± .015	.333 ± .020	.338 ± .037
	T_{pos} 2	.206 ± .044**	.326 ± .037	.324 ± .030	.332 ± .041
	T_{pos} 4	.219 ± .044**	.326 ± .029	.347 ± .043	.359 ± .050
Cross-Subject Relative Positioning	T_{pos} 1	.156 ± .024*	.290 ± .034	.283 ± .020	.288 ± .037
	T_{pos} 2	.160 ± .025**	.280 ± .031	.291 ± .026	.294 ± .044
	T_{pos} 4	.160 ± .024**	.309 ± .029	.289 ± .024	.293 ± .018
Cross-Subject Shuffling		.335 ± .020**	**.436 ± .053***	.407 ± .048	.371 ± .078
Contrastive Loss		.255 ± .029**	.269 ± .023	.285 ± .038	.339 ± .058

Note. F1-scores are presented as mean ± standard deviation. Non-pretrained SVM, RF, and GBC models are trained using handcrafted time-frequency representation and functional connectivity features, while the non-pretrained ShallowNet is trained using raw labeled EEG data. The other conditions indicate the type of pretext task utilized to pretrain the encoder to extract self-supervised learned features. The three highest obtained F1-scores are indicated in **bold**. SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier. *p-value < 0.05, **p-value <0.01

The epoch-level performance of all models, pretrained or non-pretrained, was characterized by the average F1-scores of the CV folds. Table 7 shows that no pretrained models were able to outperform the non-pretrained GBC (F1-score: $.446 \pm .052$) or the non-pretrained ShallowNet (F1-score: $.451 \pm .084$). However, the RF trained with CSS-derived SSL features achieved significantly higher F1-scores compared to the RF trained with handcrafted features (F1-score: $.436 \pm .053$ vs. $.410 \pm .033$, $p = .029$). The

baseline SVM performed poorly, with F1-scores of .091 ± .019, likely due to the high dimensionality of the handcrafted feature set, as suggested by unpublished results. All pretrained SVM models outperformed the baseline, though they still lagged behind other models. Comparing the ShallowNet models with the traditional ML models, no distinct difference is seen between the non-pretrained and pretrained conditions.

Among models trained with SSL features from different pretext tasks, CSS yielded the highest downstream F1-scores (Table 7). In contrast, the cross-subject RP task generally resulted in the lowest F1-scores. Additionally, no consistent differences were observed in F1-scores across different T_{pos} values within the within-subject and cross-subject RP tasks.

5.2 *Patient-Level Model Performance*

For the GBC trained with handcrafted features, the non-pretrained ShallowNet, and the RF trained with CSS-derived SSL features, patient-level predictions were obtained using the epoch-level predictions on the entire dataset. These epoch-level predictions were subsequently used to investigate the patient-level model performance, as well as the patient-level class-specific model performance.

Table 8: Overall and class-specific patient-level model performance in F1-score

Model	**Overall**	**ADHD**	**HC**	**MDD**	**OCD**	**SMC**
Non-pretrained GBC	.52	.58	.53	.43	.43	.64
Non-pretrained ShallowNet	.50	.46	.48	.35	.40	.82
CSS-pretrained RF	.45	.46	.55	.28	.35	.59

Note. F1-scores are obtained from patient-level predictions on the entire dataset, computed using 5-fold cross-validation. The non-pretrained GBC model is trained using handcrafted time-frequency representation and functional connectivity features, while the non-pretrained ShallowNet is trained using raw labeled EEG data. The CSS-pretrained RF is trained with SSL features learned during a CSS pretext task. Firstly, the overall F1-score of the model is presented in the 'Overall' column. Next, the class-specific F1-scores are presented in each corresponding psychiatric disorder column. RF = Random Forest; GBC = Gradient Boosting Classifier; CSS = Cross-Subject Shuffling; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

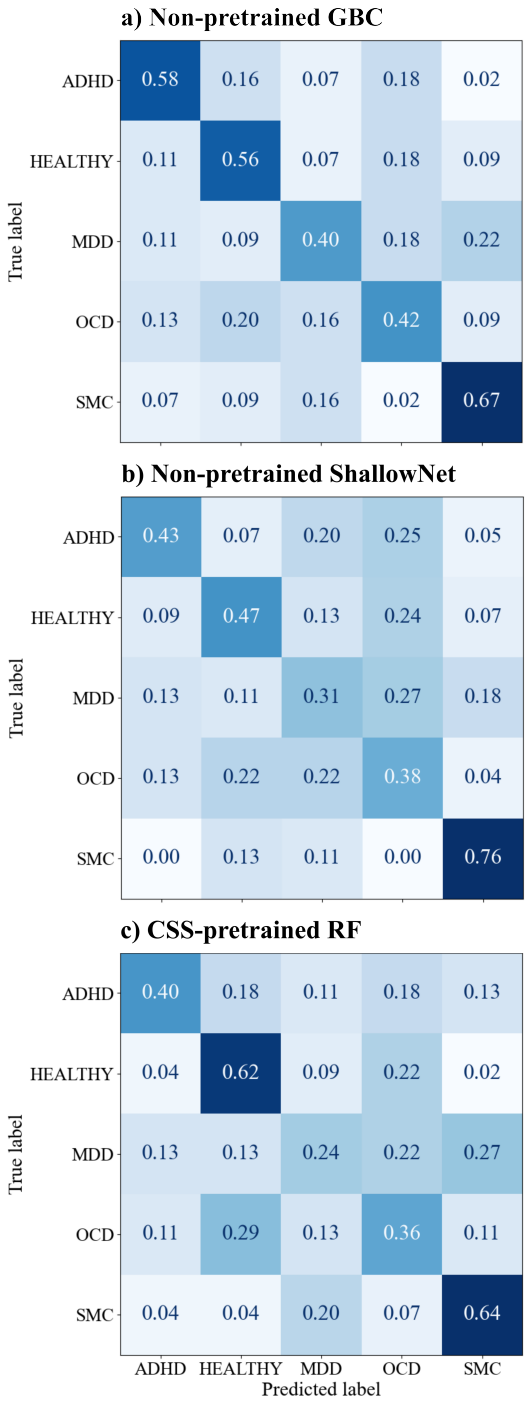

Figure 10: Normalized confusion matrices of patient-level predictions. Patient-level predictions are obtained for the entire dataset using 5-fold cross-validation. The non-pretrained GBC model (a) is trained using handcrafted time-frequency representation and functional connectivity features, while the non-pretrained ShallowNet (b) is trained using raw labeled EEG data. The CSS-pretrained RF (c) is trained with SSL features learned during a CSS pretext task. Normalized values within each confusion matrix are color-coded to represent the number of times a class was predicted divided by the number of true instances per class, with more predictions visualized as an increasingly darker blue. The range of the color gradient is determined within each individual confusion matrix. RF = Random Forest; GBC = Gradient Boosting Classifier; CSS = Cross-Subject Shuffling; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

While relatively small, a clear difference is seen in the overall F1-scores of the patient-level predictions between the three models (Table 8). Focusing on the class-specific results, in general, SMC seems to be most accurately predicted across the models, followed by ADHD and HCs (Table 8 & Figure 10). The non-pretrained models, particularly the GBC, are able to better distinguish the MDD and OCD cases from the other disorders compared to the CSS-pretrained RF. Notably, the pretrained RF did obtain the highest F1-score for the HCs, however, not much higher than the F1-score obtained by the non-pretrained GBC (0.55 vs 0.53, respectively; Table 8). The confusion matrices in Figure 10 seem to show a pattern of misclassification between MDD and SMC, and vice-versa, especially for the non-pretrained GBC and the CSS-pretrained RF.

6 DISCUSSION

The current thesis investigated SSL pretraining for EEG-based multiclass psychiatric disorder classification. Models trained with SSL features obtained with four different pretext tasks were compared with non-pretrained traditional ML models and a DNN. The aim was to reduce the limitation posed by the available labeled data and, in turn, improve upon the current EEG-based psychiatric disorder classification to provide a data-driven diagnostic alternative to the current nosology. The results show that the current implementation of the SSL-pretrained models, regardless of model type, did not outperform their non-pretrained counterparts. Nevertheless, among the SSL features tested, those derived from the CSS task yielded the best downstream performance. Moreover, the RF trained with CSS-obtained SSL features was able to outperform the respective RF baseline, however, it did not outperform the baseline GBC and ShallowNet.

These findings are inconsistent with the previous literature investigating SSL pretraining for EEG-based classification tasks (Banville et al., 2021; Mohsenvand et al., 2020; Ou et al., 2022; Xiao et al., 2024). Banville et al. (2021), for example, showed an improvement in downstream model performance when pretraining an encoder with both within- and cross-subject RP, compared to non-pretrained models. Their pretrained models held an advantage over the baseline models up to 10^3 examples per class. With more examples per class, the pretrained models were only marginally worse than the fully supervised model and still better than the model trained with handcrafted features. Moreover, Mohsenvand et al. (2020) showed similar advantages of SSL pretraining for a contrastive loss pretext task. Nevertheless, the downstream classification tasks investigated in both articles (emotion recognition, sleep staging, pathology detection) are quite different than the current multiclass classification task. The contrast

between the findings of the current thesis and the previous literature, in combination with the large difference in downstream model performance between the baseline and pretrained models, suggests that the current implementations of the pretext SSL tasks were not able to learn features that are relevant for multiclass psychiatric disorder classification.

This can be ascribed to a multitude of factors. Firstly, while we provide a comprehensive comparison between the different implementations of SSL pretraining in the current literature, none of the implementations were fully optimized. The entire pipeline has a lot of hyperparameters, from the pretext task (e.g. T_{pos}), to the pretext model (e.g. the projection head), to the encoder within the pretext model, to the downstream model. In turn, in the scope of the current thesis, it was not feasible to exhaustively tune all parameters. Rafiei et al. (2024) stress the importance of selecting the right pretext task for the respective downstream task. Our results suggest that the right pretext task, as well as the tuning of that pretext task are vital for learning SSL features that are informative for the downstream task of interest. Nonetheless, even without exhaustive tuning, the current implementation of the pretext CSS task did show promise for extracting informative SSL features for multiclass psychiatric disorder classification. Reiterating the importance of selecting the right pretext task as argued by Rafiei et al. (2024). Future research tuning the pretext tasks and models is necessary for further insights into what type of SSL pretraining is effective for multiclass psychiatric disorder classification.

Secondly, it is possible that the models had already learned all that could be learned from the labeled data, making pretraining unnecessary. This would suggest that the available participants per class was not limiting the classification performance in both the models trained with the handcrafted features and the ShallowNet trained with the raw labeled EEG data. However, previous EEG-based SSL research (Banville et al., 2021; Mohsenvand et al., 2020; Ou et al., 2022; Xiao et al., 2024) suggests that pretraining should not largely hinder performance when labeled data is not a limiting factor. Nevertheless, we do see that this expectation holds true for the CSS-pretrained RF, where the downstream model performance is only slightly worse than the general baseline performance. In turn, both explanations could be true at the same time; 1) the pretext tasks were generally not effective at learning informative features for multiclass psychiatric disorder classification, while 2) the labeled sample size was not a bottleneck for classification performance. A future investigation could perform an ablation analysis to examine the effect of the downstream sample size on the non-pretrained and SSL-pretrained downstream model performance, while iteratively limiting the labeled data further (Banville et al., 2021; Mohsenvand et al., 2020; Ou et al., 2022; Xiao et al., 2024).

While none of the SSL-derived features surpassed the performance of the non-pretrained GBC or ShallowNet, the results indicate that the CSS-based SSL features outperformed those from other pretext tasks. This finding suggests that the CSS pretext task produced the most informative features for downstream multiclass psychiatric disorder classification among the four tasks evaluated. Banville et al. (2021) previously highlighted the benefits of subject-level pretext tasks for subject-level classification. However, their study demonstrated this advantage with a cross-subject RP task, which was less effective in the current implementation. The superior performance of the CSS task in this study could stem from its independence from the temporal positioning of epochs within an original recording. Unlike the cross-subject RP task, the CSS task aligns more closely with the requirements of multiclass psychiatric disorder classification, where the position of epochs is not a relevant factor for distinguishing between disorders. This characteristic likely contributed to the CSS task's ability to extract features better suited to the downstream classification task.

Interestingly, the class-specific results of the downstream models show that the models generally performed best in classifying SMC participants, followed by ADHD participants and HCs. The etiology of SMC and ADHD are more grounded in neurobiological dysfunction than MDD and OCD (Benito-León et al., 2010; Blackburn et al., 2014; de Groot et al., 2001; Jorm et al., 2004; Reid & Maclullich, 2006; Tripp & Wickens, 2009). Conversely, HCs are characterized by the lack of neurobiological dysfunction. The clear presence or absence of neurobiological dysfunction could possibly result in more apparent features in brain activity, in turn, providing a possible explanation for the more accurate classification. An additional explanation could be the difference large differences in the age distributions between the SMC and ADHD participants compared to the other classes (Table 2). The ADHD participants, on average, were the youngest, while the SMC participant were the oldest. Older adults typically exhibit lower alpha peak frequency, higher beta power, and less parietal alpha power asymmetry compared to young adults (Laera et al., 2021), thus, even though age was not included as a predictor, the differences in age could still be encoded within the EEG data. In the future, novel implementations of normative modeling of EEG data can be investigated to control for age and other differences (Itälinna et al., 2023; Janiukstyte et al., 2023).

It is important to note several limitations of the current investigation. Firstly, a large part of the included participants were not formally diagnosed, possibly introducing noise in the data. To this point, the number of formally diagnosed participants was not evenly distributed among the classes (Table 2). This could result in differences in the amount of heterogeneity and noise per class, in turn, introducing bias in the class-specific

results. With an effective SSL pipeline, future studies could permit to include a smaller, but true, sample of psychiatric patients for the development of ML models, thus providing more robust insights.

Secondly, while the current approach utilizes nested CV for model development and 5-fold CV for obtaining the predictions on the dataset, no separate hold-out set was available for a final generalization error analysis. Unfortunately, the replication set provided with the TDBRAIN dataset, was not suitable for external validation of our multiclass classification task. If in the future the TDBRAIN dataset is updated with additional participants, an external validation for the generalizability of the models can still be performed. Nevertheless, even without the external validation, we argue that by evaluating our models on multiple held-out test sets during the nested CV, the current approach still provides a robust measure of the generalization error (Varma & Simon, 2006).

Finally, while the data-driven approach employed in this study offers an objective diagnostic framework, challenges persist due to the substantial within-disorder heterogeneity and between-disorder homogeneity inherent in traditional psychiatric classifications defined by symptom complexes with ambiguous boundaries (Feczko et al., 2019; Fu et al., 2019; McTeague & Lang, 2012). We propose that future research should consider combining objective data-driven methods with dimensional diagnostic frameworks, such as the Hierarchical Taxonomy of PsychoPathology (HiTOP) method (Kotov et al., 2017). Integrating these approaches could leverage the strengths of data-driven ML for biomarker-based diagnosis while addressing the limitations of traditional nosology by reducing within-disorder heterogeneity and between-disorder homogeneity through dimensional classification.

Nonetheless, the current thesis additionally has multiple strong points. Firstly, to the extend of our knowledge, we provide the first comparison of the included pretext SSL tasks for EEG-based classification, and the first investigation of any pretext SSL tasks for multiclass psychiatric disorder classification. Within this investigation we include a multitude of different models ranging in complexity, trained with a range of different inputs; handcrafted features, the raw EEG data, or the SSL features. In addition, during the entire pipeline, data leakage has been prevented by using a stratified grouped K-fold splitting approach for all models, in combination with nested CV for the development of all downstream models. Thus, the current thesis provides a robust and unbiased estimation of the model performances. While the SSL pretraining did not increase downstream classification performance, we provide a foundation for future investigations for EEG-based multiclass psychiatric disorder classification.

Building upon this foundation with the proposed future directions, we believe the development of a data-driven diagnostic framework for psychiatric patients is possible. This framework would enable objective and precise diagnoses of psychiatric disorders with well-defined boundaries, ultimately facilitating personalized and effective treatment for each individual and facilitating research into the underlying pathophysiological mechanisms. Moreover, the described SSL pretraining framework can have great implications for other EEG-based classification tasks, such as Alzheimer's disease detection (Waskom, 2021), or any other classification task where the labeled data is scarce compared to the amount of unlabeled data. This scenario is a common challenge across various types of medical data (Dercksen et al., 2019; Schäfer et al., 2024; Sufi, 2024). Thus, the current investigation offers perspective and hope for any patient population that could benefit from improved classification through SSL pretraining.

7 CONCLUSION

In conclusion, the current thesis provides a comprehensive comparison of the current EEG-based pretext SSL tasks for multiclass psychiatric disorder classification. In contrast to our expectation, the current SSL pretraining implementation did not yield better model performance than the non-pretrained models. However, the CSS pretext SSL task did show some ability to obtain informative features, stressing the need to select and tune the right pretext task for the corresponding downstream multiclass classification. Examining the patient-level predictions of the models offered valuable insights into the classification of specific psychiatric disorders. Notably, the models were generally able to distinguish the SMC, ADHD and HC cases. These results imply that the the features in the data, either as raw input, handcrafted features, or SSL features, are more distinct for these disorders compared to OCD and SMC. Lastly, no clear distinction was seen in the effect of SSL pretraining between the traditional ML models and the DNN. Overall, these findings offer valuable insight into the potential of SSL pretraining for EEG-based multiclass psychiatric disorder classification, providing the foundation for an objective data-driven diagnostic framework for psychiatric patients.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowl-*

edge Discovery & Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

Association, A. P. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th).

Babadi, B., & Brown, E. N. (2014). A Review of Multitaper Spectral Analysis [Conference Name: IEEE Transactions on Biomedical Engineering]. *IEEE Transactions on Biomedical Engineering*, *61*(5), 1555–1564. https://doi.org/10.1109/TBME.2014.2311996

Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.-A., & Gramfort, A. (2021). Uncovering the structure of clinical EEG signals with self-supervised learning [Publisher: IOP Publishing]. *Journal of Neural Engineering*, *18*(4), 046020. https://doi.org/10.1088/1741-2552/abca18

Barry, R. J., Clarke, A. R., Johnstone, S. J., Magee, C. A., & Rushby, J. A. (2007). EEG differences between eyes-closed and eyes-open resting conditions. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *118*(12), 2765–2773. https://doi.org/10.1016/j.clinph.2007.07.028

Benito-León, J., Mitchell, A. J., Vega, S., & Bermejo-Pareja, F. (2010). A population-based study of cognitive function in older people with subjective memory complaints. *Journal of Alzheimer's disease: JAD*, *22*(1), 159–170. https://doi.org/10.3233/JAD-2010-100972

Blackburn, D. J., Wakefield, S., Shanks, M. F., Harkness, K., Reuber, M., & Venneri, A. (2014). Memory difficulties are not always a sign of incipient dementia: A review of the possible causes of loss of memory efficiency. *British Medical Bulletin*, *112*(1), 71–81. https://doi.org/10.1093/bmb/ldu029

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, June). A Simple Framework for Contrastive Learning of Visual Representations [arXiv:2002.05709 [cs, stat]]. https://doi.org/10.48550/arXiv.2002.05709

Cieslak, M. C., Castelfranco, A. M., Roncalli, V., Lenz, P. H., & Hartline, D. K. (2020). T-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis. *Marine Genomics*, *51*, 100723. https://doi.org/10.1016/j.margen.2019.100723

Cohen, M. X. (2019). A better way to define and describe Morlet wavelets for time-frequency analysis. *NeuroImage*, *199*, 81–86. https://doi.org/10.1016/j.neuroimage.2019.05.048

de Groot, J. C., de Leeuw, F. E., Oudkerk, M., Hofman, A., Jolles, J., & Breteler, M. M. (2001). Cerebral white matter lesions and subjective cognitive dysfunction: The Rotterdam Scan Study. *Neurology*, *56*(11), 1539–1545. https://doi.org/10.1212/wnl.56.11.1539

Dercksen, K., Bulten, W., & Litjens, G. (2019, May). Dealing with Label Scarcity in Computational Pathology: A Use Case in Prostate Cancer Classification [arXiv:1905.06820]. https://doi.org/10.48550/arXiv.1905.06820

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Emre, İ. E., Erol, Ç., Taş, C., & Tarhan, N. (2023). Multi-class classification model for psychiatric disorder discrimination. *International Journal of Medical Informatics*, *170*, 104926. https://doi.org/10.1016/j.ijmedinf.2022.104926

Falcon, W., Borovec, J., Wälchli, A., Eggert, N., Schock, J., Jordan, J., Skafte, N., Ir1dXD, Bereznyuk, V., Harris, E., Murrell, T., Yu, P., Præsius, S., Addair, T., Zhong, J., Lipin, D., Uchida, S., Bapat, S., Schröter, H., … Bakhtin, A. (2020, May). PyTorchLightning/pytorch-lightning: 0.7.6 release. https://doi.org/10.5281/zenodo.3828935

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., & Fair, D. A. (2019). The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes [Publisher: Elsevier]. *Trends in Cognitive Sciences*, *23*(7), 584–601. https://doi.org/10.1016/j.tics.2019.03.009

Fu, C. H. Y., Fan, Y., & Davatzikos, C. (2019). Addressing heterogeneity (and homogeneity) in treatment mechanisms in depression and the potential to develop diagnostic and predictive biomarkers. *NeuroImage: Clinical*, *24*, 101997. https://doi.org/10.1016/j.nicl.2019.101997

Fuchs, T. (2010). Subjectivity and Intersubjectivity in Psychiatric Diagnosis. *Psychopathology*, *43*(4), 268–274. https://doi.org/10.1159/000315126

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python [Publisher: Frontiers]. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*(4), 468–484. https://doi.org/10.1016/0013-4694(83)90135-9

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R.,

Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy [Publisher: Nature Publishing Group]. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hilbert, D. (1912). *Grundzüge einer allgemeinen theorie der linearen integralgleichungen*. Leipzig, B. G. Teubner. Retrieved June 14, 2024, from http://archive.org/details/grundzugeallgoohilbrich

Homola, D. (2015, May). BorutaPy. Retrieved June 14, 2024, from https://danielhomola.com/feature%20selection/phd/borutapy-an-all-relevant-feature-selection-method/

Hunter, J. D. (n.d.). Matplotlib: A 2D Graphics Environment | IEEE Journals & Magazine | IEEE Xplore. Retrieved December 2, 2024, from https://ieeexplore-ieee-org.ru.idm.oclc.org/document/4160265

Huynh, N., Yan, D., Ma, Y., Wu, S., Long, C., Sami, M. T., Almudaifer, A., Jiang, Z., Chen, H., Dretsch, M. N., Denney, T. S., Deshpande, R., & Deshpande, G. (2024). The Use of Generative Adversarial Network and Graph Convolution Network for Neuroimaging-Based Diagnostic Classification. *Brain Sciences*, *14*(5), 456. https://doi.org/10.3390/brainsci14050456

Itälinna, V., Kaltiainen, H., Forss, N., Liljeström, M., & Parkkonen, L. (2023). Using normative modeling and machine learning for detecting mild traumatic brain injury from magnetoencephalography data [Publisher: Public Library of Science]. *PLOS Computational Biology*, *19*(11), e1011613. https://doi.org/10.1371/journal.pcbi.1011613

Janiukstyte, V., Owen, T. W., Chaudhary, U. J., Diehl, B., Lemieux, L., Duncan, J. S., de Tisi, J., Wang, Y., & Taylor, P. N. (2023). Normative brain mapping using scalp EEG and potential clinical application [Publisher: Nature Publishing Group]. *Scientific Reports*, *13*(1), 13442. https://doi.org/10.1038/s41598-023-39700-7

Jing, L., & Tian, Y. (2021). Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey [Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(11), 4037–4058. https://doi.org/10.1109/TPAMI.2020.2992393

Jorm, A. F., Butterworth, P., Anstey, K. J., Christensen, H., Easteal, S., Maller, J., Mather, K. A., Turakulov, R. I., Wen, W., & Sachdev, P. (2004). Memory complaints in a community sample aged 60-64 years: Associations with cognitive functioning, psychiatric symptoms, medical conditions, APOE genotype, hippocampus and amygdala volumes, and white-matter hyperintensities. *Psychological Medicine*, *34*(8), 1495–1506. https://doi.org/10.1017/s0033291704003162

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., … Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*, 1–13. https://doi.org/10.18637/jss.v036.i11

Laera, G., Arcara, G., Gajewski, P. D., Kliegel, M., & Hering, A. (2021). Age-related modulation of EEG time-frequency responses in prospective memory retrieval. *Neuropsychologia*, *155*, 107818. https://doi.org/10.1016/j.neuropsychologia.2021.107818

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces [arXiv:1611.08024 [cs, q-bio, stat]]. *Journal of Neural Engineering*, *15*(5), 056013. https://doi.org/10.1088/1741-2552/aace8c

Li, L., Xiao, L., & Chen, L. (2009). Differences of EEG between Eyes-Open and Eyes-Closed States Based on Autoregressive Method. *Journal of Electronic Science and Technology*, *7*(2), 175–179. Retrieved June 14, 2024, from https://www.journal.uestc.edu.cn/en/article/id/1644

Liu, B., Chang, H., Peng, K., & Wang, X. (2022). An End-to-End Depression Recognition Method Based on EEGNet. *Frontiers in Psychiatry*, *13*, 864393. https://doi.org/10.3389/fpsyt.2022.864393

Louis, E. K. S., Frey, L. C., Britton, J. W., Frey, L. C., Hopp, J. L., Korb, P., Koubeissi, M. Z., Lievens, W. E., Pestana-Knight, E. M., & Louis, E. K. S. (2016a). Appendix 1. The Scientific Basis of EEG: Neurophysiology of EEG Generation in the Brain. In *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants [Internet]*. American Epilepsy Society. Retrieved June 3, 2024, from https://www.ncbi.nlm.nih.gov/books/NBK390351/

Louis, E. K. S., Frey, L. C., Britton, J. W., Frey, L. C., Hopp, J. L., Korb, P., Koubeissi, M. Z., Lievens, W. E., Pestana-Knight, E. M., & Louis, E. K. S. (2016b). Introduction. In *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants [Internet]*. American Epilepsy Society. Retrieved June 3, 2024, from https://www.ncbi.nlm.nih.gov/books/NBK390346/

Maaten, L. v. d., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. Retrieved June 16, 2024, from http://jmlr.org/papers/v9/vandermaaten08a.html

Mari, T., Henderson, J., Maden, M., Nevitt, S., Duarte, R., & Fallon, N. (2022). Systematic Review of the Effectiveness of Machine Learning Algorithms for Classifying Pain Intensity, Phenotype or Treatment Outcomes Using Electroencephalogram Data [Publisher: Elsevier]. *The Journal of Pain*, *23*(3), 349–369. https://doi.org/10.1016/j.jpain.2021.07.011

McTeague, L. M., & Lang, P. J. (2012). The Anxiety Spectrum and the Reflex Physiology of Defense: From Circumscribed Fear to Broad Distress [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.21891]. *Depression and Anxiety*, *29*(4), 264–281. https://doi.org/10.1002/da.21891

Michelini, G., Palumbo, I. M., DeYoung, C. G., Latzman, R. D., & Kotov, R. (2021). Linking RDoC and HiTOP: A new interface for advancing psychiatric nosology and neuroscience. *Clinical Psychology Review*, *86*, 102025. https://doi.org/10.1016/j.cpr.2021.102025

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, September). Efficient Estimation of Word Representations in Vector Space [arXiv:1301.3781 [cs]]. https://doi.org/10.48550/arXiv.1301.3781

Mohsenvand, M. N., Izadi, M. R., & Maes, P. (2020). Contrastive Representation Learning for Electroencephalogram Classification [ISSN: 2640-3498]. *Proceedings of the Machine Learning for Health NeurIPS Workshop*, 238–253. Retrieved August 29, 2024, from https://proceedings.mlr.press/v136/mohsenvand20a.html

Mumtaz, W., Ali, S. S. A., Yasin, M. A. M., & Malik, A. S. (2018). A machine learning framework involving EEG-based functional connectivity to diagnose major depressive disorder (MDD). *Medical & Biological Engineering & Computing*, *56*(2), 233–246. https://doi.org/10.1007/s11517-017-1685-z

Organization, W. H. (1992). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines* [Google-Books-ID: DFMoDgAAQBAJ].

Ou, Y., Sun, S., Gan, H., Zhou, R., Yang, Z., Ou, Y., Sun, S., Gan, H., Zhou, R., & Yang, Z. (2022). An improved self-supervised learning for EEG classification [Cc_license_type: cc_by Number: mbe-19-07-325 Primary_atype: Mathematical Biosciences and Engineering Subject_term: Research article Subject_term_id: Research article]. *Mathematical Biosciences and Engineering*, *19*(7), 6907–6922. https://doi.org/10.3934/mbe.2022325

pandas development team, T. (2024, September). Pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.13819579

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019, December). PyTorch: An Imperative Style, High-Performance Deep Learning Library [arXiv:1912.01703]. https://doi.org/10.48550/arXiv.1912.01703

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. Retrieved June 10, 2024, from http://jmlr.org/papers/v12/pedregosa11a.html

Pup, F. D., Zanola, A., Tshimanga, L. F., Mazzon, P. E., & Atzori, M. (2024). SelfEEG: A Python library for Self-Supervised Learning in Electroencephalography. *Journal of Open Source Software*, *9*(95), 6224. https://doi.org/10.21105/joss.06224

Rafiei, M. H., Gauthier, L. V., Adeli, H., & Takabi, D. (2024). Self-Supervised Learning for Electroencephalography [Conference Name: IEEE Transactions on Neural Networks and Learning Systems]. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(2), 1457–1471. https://doi.org/10.1109/TNNLS.2022.3190448

Reid, L. M., & Maclullich, A. M. J. (2006). Subjective memory complaints and cognitive impairment in older people. *Dementia and Geriatric Cognitive Disorders*, *22*(5-6), 471–485. https://doi.org/10.1159/000096295

Saeidi, M., Karwowski, W., Farahani, F. V., Fiok, K., Taiar, R., Hancock, P. A., & Al-Juaid, A. (2021). Neural Decoding of EEG Signals with Machine Learning: A Systematic Review. *Brain Sciences*, *11*(11), 1525. https://doi.org/10.3390/brainsci11111525

Schäfer, R., Nicke, T., Höfener, H., Lange, A., Merhof, D., Feuerhake, F., Schulz, V., Lotz, J., & Kiessling, F. (2024). Overcoming data scarcity in biomedical imaging with a foundational multi-task model [Publisher: Nature Publishing Group]. *Nature Computational Science*, *4*(7), 495–509. https://doi.org/10.1038/s43588-024-00662-z

Sufi, F. (2024). Addressing Data Scarcity in the Medical Domain: A GPT-Based Approach for Synthetic Data Generation and Feature Extraction [Number: 5 Publisher: Multidisciplinary Digital Publishing Institute]. *Information*, *15*(5), 264. https://doi.org/10.3390/info15050264

Taskesen, E., & Reinders, M. J. T. (2016). 2D Representation of Transcriptomes by t-SNE Exposes Relatedness between Human Tissues [Publisher: Public Library of Science]. *PLOS ONE*, *11*(2), e0149853. https://doi.org/10.1371/journal.pone.0149853

Tripp, G., & Wickens, J. R. (2009). Neurobiology of ADHD. *Neuropharmacology*, *57*(7), 579–589. https://doi.org/10.1016/j.neuropharm.2009.07.026

van Dijk, H., van Wingen, G., Denys, D., Olbrich, S., van Ruth, R., & Arns, M. (2022). The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database [Publisher: Nature Publishing Group]. *Scientific Data*, *9*(1), 333. https://doi.org/10.1038/s41597-022-01409-z

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*, 91. https://doi.org/10.1186/1471-2105-7-91

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python [Publisher: Nature Publishing Group]. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Xiao, T., Wang, Z., Zhang, Y., lv, H., Wang, S., Feng, H., & Zhao, Y. (2024). Self-supervised Learning with Attention Mechanism for EEG-based seizure detection. *Biomedical Signal Processing and Control*, *87*, 105464. https://doi.org/10.1016/j.bspc.2023.105464

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019, February). How Powerful are Graph Neural Networks? [arXiv:1810.00826 [cs, stat]]. Retrieved June 3, 2024, from http://arxiv.org/abs/1810.00826

Zhou, Y., Huo, H., Hou, Z., & Bu, F. (2023). A deep graph convolutional neural network architecture for graph classification. *PLOS ONE*, *18*(3), e0279604. https://doi.org/10.1371/journal.pone.0279604

APPENDIX A: PYTHON PACKAGES

Table 9: Package Utilization

Package	Utilization	Source
MNE (version 1.6)	EEG preprocessing, handcrafted features	(Gramfort et al., 2013)
BorutaPy (version 0.3)	feature selection	(Homola, 2015)
Scikit-learn (version 1.4)	traditional ML model development	(Pedregosa et al., 2011)
Pytorch Lightning (version 2.2)	neural network development	(Falcon et al., 2020)
PyTorch (version 2.2)	neural network development	(Paszke et al., 2019)
Optuna (version 3.6)	neural network hyperparameter tuning	(Akiba et al., 2019)
NumPy (version 1.26)	general	(Harris et al., 2020)
pandas (version 2.2)	general	(pandas development team, 2024)
SciPy (version 1.13)	general	(Virtanen et al., 2020)
Matplotlib (version 3.8)	plotting	(Hunter, n.d.)
Seaborn (version 0.13)	plotting	(Waskom, 2021)
joblib (version 1.4)	general	-
yaml (version 0.2.5)	general	-

APPENDIX B: FEATURE SELECTION

Table 10: Boruta-Selected Feature Types and Counts

Feature type	EC
FC features:	
Total	180
Statistical features:	
Mean	45
Standard deviation	45
Median	45
Skewness	3
Kurtosis	1
Total	139
Frequency bands:	
Delta	63
Theta	63
Alpha	63
Beta	63
Gamma	67
Electrode groups:	
L-frontal	55
M-frontal	55
R-frontal	55
L-central	55
M-central	55
R-central	55
L-posterior	57
M-posterior	56
R-posterior	56

Note. FC = functional connectivity; EC = eyes-closed; L = left; M = mid; R = right.

APPENDIX C: CLUSTER ANALYSIS EXAMPLES

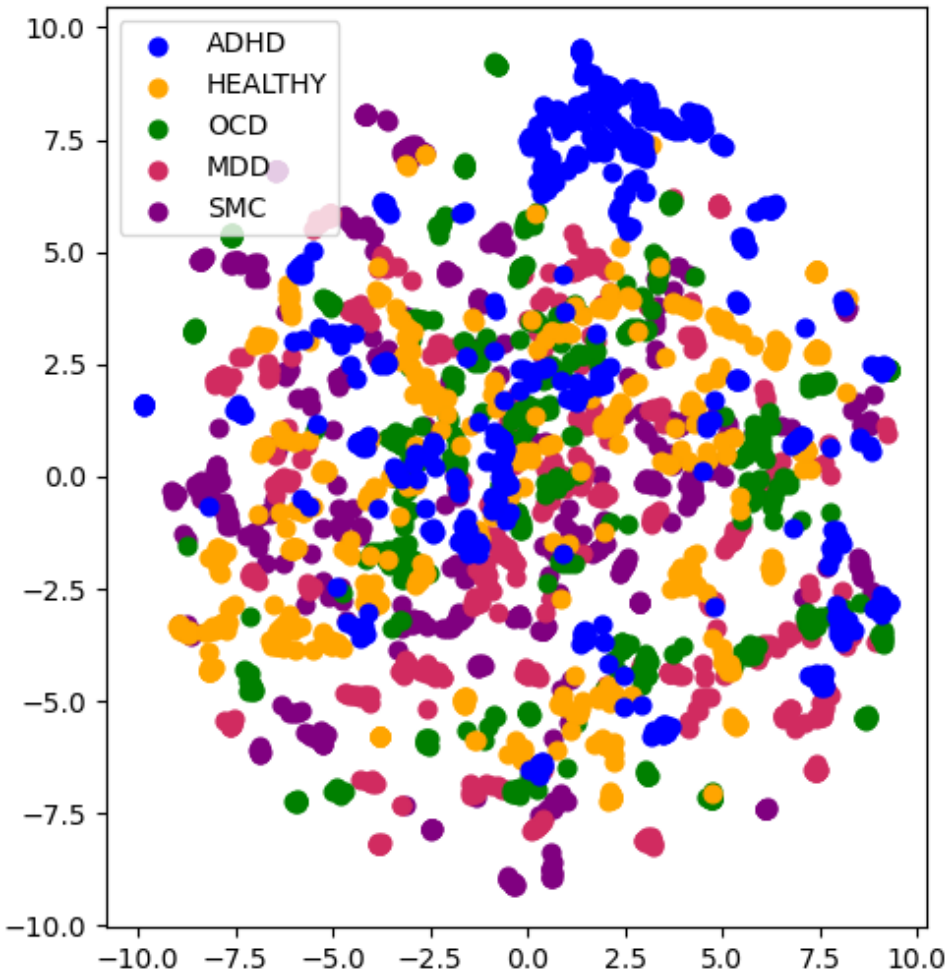

Figure 11: Representative example of t-SNE visualization for all feature sets. This particular example is derived from the handcrafted features.