Resting-State EEG Biomarkers for Predicting Psychiatric Disorders Using Machine Learning

Tuur Smolders[1], s1018330, Master Medical Biology Internship report
Supervisor: Dr. Marijn van Wingerden[2]

[1]Master Medical Biology, specialization Neurobiology, Radboud University, Nijmegen, The Netherlands. Master Data Science & Society, specialization Health, Tilburg University, Tilburg, The Netherlands. [2]Department of Cognitive Science and Artificial Intelligence, Tilburg School of Humanities and Digital Sciences, Tilburg University, Tilburg, The Netherlands.

Background: Discovery of physiological biomarkers for psychiatric disorders are vital for addressing heterogeneity in current nosological diagnostic methodology, aiming to advance towards objective diagnoses and patient-specific psychiatric interventions.

Aim: The current study investigates the potential utility of electroencephalogram-derived statistical time-frequency features and functional connectivity (FC) features of brain activity as biomarkers for classifying psychiatric disorders, providing insights into their underlying patterns of psychopathology.

Method: The study evaluated the predictive performance of FC features and statistical features extracted from time-frequency representations (TFR) of eyes-closed (EC) and eyes-open (EO) resting-state EEG recordings. Participants included those with attention-deficit hyperactivity disorder (ADHD), major depressive disorder (MDD), obsessive-compulsive disorder (OCD), subjective memory complaints (SMC), and healthy controls (HCs). Multiclass standard machine learning (ML) models and novel Graph Convolutional Network (GCN) models, including graphLambda, were trained and evaluated.

Results: Random Forest (RF) and Gradient Boosting Classifier (GBC) models trained with the statistical TFR features along with the FC features achieved higher F1-scores than the baseline RF and GBC models trained with exclusively the statistical TFR features (0.411, 0.440, 0.361, 0.370, respectively). Generally, models trained with EC- and EO-derived feature sets outperformed the models trained with feature sets derived from the ratio between the EC and EO EEG recording.

Furthermore, analysis of class predictions indicated that ADHD and SMC were most distinguishable from other classes. Lastly, the GCN models were unable to outperform the standard ML algorithms.

Conclusion: FC patterns in brain activity represent important biomarkers for certain psychiatric disorders. For the included psychiatric disorders in the current study, EC and EO resting-state brain activity yield better biomarkers than the difference between EC and EO resting-state brain activity. Lastly, these biomarkers in brain activity are most distinct for ADHD and SMC comparison to the other classes.

Abbreviations

ADHD, attention-deficit hyperactivity disorder; ASD, autism spectrum disorder; BN layer, batch normalization layer; CNN, convolutional neural network; CV, cross-validation; DO layer, dropout layer; DPSS, discrete prolate Slepian tapers; DSM, Diagnostic and Statistical Manual of Mental Disorders; EC, eyes-closed; ECG, electrocardiogram; EEG, electroencephalogram; EMG, electromyogram; EO, eyes-open; FC layer, fully connected layer; FFT, fast Fourier transform; fMRI, functional magnetic resonance imaging; GAT, graph attention network; GBC, gradient boosting classifier; GCN, graph convolutional network; GIN, graph isomorphism network; HC, healthy control; HiTOP, Hierarchical Taxonomy of Psychopathology; ICD, International Classification of Disorders; MDD, major depressive disorder; ML, machine learning; NLL, negative log likelihood; OCD, obsessive-compulsive disorder; OOB, out-of-bag; PSD, power spectral density; RDoC, Research Domain Criteria; RF, random forest; SMC, subjective memory complaints; SVM, support vector machine; TDBRAIN, Two Decades-Brainclinic Research Archive for Insights in Neurophysiology; TFR, time-frequency representation; t-SNE, t-distributed stochastic neighbor embedding

1 Introduction

In current clinical practice, psychiatric disorders are diagnosed according to the International Classification of Disorders (ICD; World Health Organization, 1992) and the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2013). The ICD and DSM categorically diagnose disorders based on signs and symptoms observed by a clinician. This provides a relatively simple diagnostic methodology for complex pathologies, in turn, providing a way forward for treatment. The nosological nature of the current diagnostic methodology, however, is limited by the subjectivity between the clinicians and the presented complaints by the patient (Fuchs, 2010). Furthermore, due to the high heterogeneity within psychiatric disorders, and the high homogeneity among psychiatric disorders, it is difficult to accurately diagnose a patient on clinical symptoms alone (Feczko et al., 2019; Fu et al., 2019; McTeague & Lang, 2012). In addition, besides the issues with diagnostic classification, there is a lack of a complete understanding of the etiology of most mental disorders, making it difficult to address the heterogeneity in nosological diagnosis, and to develop better treatments (Michelini et al., 2021). Concurrently, advancements in the understanding of the etiology of psychiatric disorders are hindered by common misdiagnosis and high heterogeneity.

To address these shortcomings, alternatives to the traditional nosology, such as the Hierarchical Taxonomy of Psychopathology (HiTOP) method (Kotov et al., 2017) and the Research Domain Criteria (RDoC) framework (Cuthbert & Insel, 2013; Kozak & Cuthbert, 2016), have been proposed. The HiTOP model, for example, formulates psychopathological syndromes and their respective subtypes based on the observed covariation of symptoms, thereby clustering related symptoms and minimizing heterogeneity. In addition, it maps out comorbidity by combining co-occurring syndromes in spectra, and it addresses the unclear boundaries between disorders by characterizing the observed phenomena dimensionally. Moreover, in contrast to the traditional view of disorders as symptom complexes, the RDoC framework encourages research on the fundamental primary behavioral functions (e.g. cognition) and the neural systems responsible for these functions, to further the understanding of the transdiagnostic biobehavioral systems underlying psychopathology (Cuthbert & Insel, 2013; Kozak & Cuthbert, 2016; Michelini et al.,

2021). The RDoC framework, therefore, provides a research classification system rooted in neurobiology.

In addition to the alternative nosologies, current diagnostic methodology can greatly improve with data-driven approaches. For example, Hofman et al. (2023) performed a data-driven symptom-based unsupervised clustering analysis of psychiatric patients. Using this data-driven approach, they were able to uncover multiple psychiatric phenotypes within a population-based sample of participants with depressive disorders, anxiety disorders, and sleep-wake disorders, in turn, providing insights into patterns of symptom co-occurrence. Another advantage of data-driven approaches over the current nosology is the use of measurable neurobiological biomarkers, such as features in brain activity, for the classification of psychiatric disorders (Alnagger et al., 2023; Calhoun et al., 2021; Zhang et al., 2020). Here, objective physiological measurements are used to diagnose patients, as opposed to the clinical observations of a clinician. In turn, these data-driven approaches are able to provide insight into the mechanistic understanding of psychiatry, identify dimensional subtypes, and can predict individual treatment response. Thus, biomarker-based data-driven approaches, encompassing the psychopathology of the psychiatric disorders, is needed to improve the precision of diagnoses and provide personalized treatment.

There is large interest in the use of electroencephalography (EEG) to measure brain activity as a biomarker for psychiatric disorders. EEG measures the electrical activity in the brain through multiple electrodes placed on the scalp (Louis et al., 2016b). Subsequently, the electrical activity measured by EEG recordings is generated by similarly oriented groups of cerebral cortical neurons near the scalp (Louis et al., 2016a). This grouping of neuronal activity and the measured voltages first needing to pass through multiple biological filters, such as the skull, leads to a poor spatial resolution. This does, however, not influence the excellent temporal resolution of EEG. Furthermore, one of the biggest advantages of EEG is that it is non-invasive and cost-effective, as compared to other methodologies for measuring brain activity (Louis et al., 2016a, 2016b). EEG, therefore, could provide a relatively easily available and objective measure for a data-driven approach to diagnose psychiatric disorders, encompassing the psychopathology of the disorder.

A multitude of different features can be extracted from an EEG recording. Moreover, EEG recordings themselves can be obtained under different conditions, leading to different features. For example, spontaneous brain activity can be recorded with EEG by recording the participant while they are resting with either their eyes open or closed, called resting-state EEG data (Barry et al., 2007; Ebrahimzadeh et al., 2022). Additionally, evoked brain activity can be recorded with EEG during the execution of a task, called task-based EEG data (Ebrahimzadeh et al., 2022; Karamzadeh et al., 2013).

Brain activity features that can be extracted from these recordings, include, for example, the Power Spectral Density (PSD) of the signal, encoding the distribution of power over frequency (Stoica & Moses, 2005). Angulo-Ruiz et al. (2023), for instance, investigated the difference in resting-state EEG dynamics in children with Autism Spectrum Disorder (ASD), and children with normal development. This was done by extracting, among other features, PSD estimates from the EEG recordings. They find that children with ASD exhibit increased power in the high frequencies (beta and gamma) of the PSD. Moreover, Amico et al. (2023) investigated differences in PSD between Major Depressive Disorder (MDD) patients with suicidal ideation and healthy controls (HCs). Their results, however, did not suggest any frequency-specific anomalies in MDD patients, compared to HCs. In addition to PSD estimates from a singular EEG recording, ratios between PSD estimates from different EEG recordings can serve as features of brain activity as well. For example, Fonseca et al. (2013) have shown that the difference in alpha power between eyes-closed (EC) and eyes-open (EO) resting-state conditions, also called alpha reactivity, was reduced in children with ADHD compared to HCs.

Nevertheless, the PSD of a signal is limited as a feature of brain activity, in that it lacks valuable temporal information. The Time-Frequency representation (TFR) of a signal, conversely, encodes the time and frequency domains simultaneously (L. Cohen, 1995; Sejdic et al., 2009). Thus, features extracted from the TFR can entail changes in power per frequency over time. Suzuki et al. (2023), for example, performed time-frequency analyses on EEG recordings of individuals with obsessive-compulsive disorder (OCD). During these recordings, participants' responses were monitored after the commission of errors. By investigating the TFR of these recordings, they showed that individuals with OCD exhibit abnormally enhanced

midfrontal theta total power on committing an error, compared to controls. However, a limitation of the measured power in a TFR and PSD is that the measured absolute power of frequencies are heavily dependent on person-specific and electrode-specific conductivity, for example, due to differences in gain between electrodes, or differences in head tissue conductivity between individuals (Akalin Acar et al., 2016).

Another feature that can be extracted from EEG, which is more robust against conductivity differences, is the functional connectivity (FC) between different brain regions, for example by measuring the synchrony of the phase of signals from two different electrodes, or electrode groups (Aviyente et al., 2017; Morales & Bowers, 2022). Furlong et al. (2021), for instance, analyzed the FC within the brain activity of children with attention-deficit hyperactivity disorder (ADHD), and typically developing children. FC was measured as the synchronization of the time series of each pair of electrodes. Their findings suggest that elevated ADHD symptom severity within the ADHD group is associated with a hyper-connected neural network. The authors did not observe any differences in FC between the ADHD and control groups, however, other literature has reported decreased global connectivity, and increased local connectivity, in individuals with ADHD compared to controls (Beare et al., 2017; Wang et al., 2008). Moreover, FC biomarkers have been shown to be associated with other disorders than ADHD as well, for example, FC biomarkers have been identified for patients with MDD (M. D. Greicius et al., 2007; Helm et al., 2018; Kaiser et al., 2015), OCD (Geffen et al., 2022; Gürsel et al., 2018; Peng et al., 2022), and for patients with memory loss disorders, such as Alzheimer's disease (M. Greicius, 2008; Iorio-Morin et al., 2022, p. 7; van den Heuvel & Hulshoff Pol, 2010).

In the last decade, research into the aforementioned EEG-derived features as biomarkers for the diagnosis of psychiatric disorders has been vast, especially with the ever-increasing possibilities of data-driven machine learning (ML) approaches. For example, Mumtaz et al. (2017), used classification models such as support vector machine (SVM), logistic regression and Naïve Bayesian to discriminate between MDD patients and HCs. These models were trained with EEG-derived FC features, resulting in accuracies up to 98%. In addition, Emre et al. (2023) have developed multiclass classification models, such as random forest (RF) and SVM, classifying bipolar disorder, ADHD, depression, OCD, opioid addiction, posttraumatic stress disorder (PTSD), schizophrenia, and HCs. Their multiclass classification

models were trained with EEG-derived absolute power values belonging to four frequency bands (alpha, beta, delta, theta), resulting in accuracies up to 84%.

While classical machine learning models have been proven to decently distinguish between different psychiatric disorders using EEG-derived features, lately, deep learning approaches, such as graph convolutional networks (GCNs), have been suggested to better utilize the intrinsic FC patterns in the brain for the classification of psychiatric disorders (Huynh et al., 2024; Xu et al., 2019; Yin et al., 2023; Zhou et al., 2023). The utilization of the intrinsic FC patterns in the brain is possible, by using the FC features as input for a spatially-informed GCN, which in turn computes spatially-informed feature maps. Due to the convolutional layers, GCNs are able to extract patterns from the data using multidimensional convolutional filters. These extracted patterns are simple at first, however, with additional convolutional layers they become more complex. The convolution layers, therefore, are able to extract their own feature maps from the original input, while concurrently reducing the dimensionality. This is unlike classical learning procedures, which often need separate feature selection techniques to reduce the dimensionality and filter out uninformative features. The difference between convolutional layers in convolutional neural networks (CNNs) and GCNs, is that in a CNN it is assumed that direct neighbors of a point in a grid hold valuable information for convolution, however, in a GCN this assumption is more relaxed and information can originate from relational connections with nodes anywhere in a graph. Thus, GCNs are able to exploit the relationships between entities of graph-structure data, such as networks of EEG sensors. In the case of EEG data, electrodes are represented as nodes within a graph, and the possible connection between the nodes as edges (Bullmore & Sporns, 2009; Furlong et al., 2021). FC information of these edges can either be binary encoded, by the exclusion or inclusion of an edge based on, for example, a minimum threshold of synchrony, or it can be continuously encoded if the GCN architecture allows for the input of edge attributes. Huynh et al. (2024), for example, have utilized GCNs to distinguish between ADHD and HCs with FC features obtained from resting-state functional magnetic resonance imaging (fMRI), resulting in accuracies up to 75% with a GCN model. A similar study was performed by Yin et al. (2023), in which they proposed a schizophrenia automatic recognition model based on a GCN architecture. This model was trained with resting-state EEG-

derived features, resulting in accuracies up to 90%. Building on recent advancements in GCNs, graphLambda, an even more advanced deep learning architecture, called a Fusion Graph Neural Network, could potentially further improve classification performance by combining the strengths of a GCN architecture with other neural network architectures called Graph Attention Network (GAT), and Graph Isomorphism Network (GIN). In short, GATs enable the specification of different weights to different nodes in a neighborhood (Veličković et al., 2018), and GINs achieve maximum discriminative power among graph neural networks (Xu et al., 2019). Another advantage of the GAT architecture, is that it can be modified to allow for the inclusion of multi-dimensional edge attributes in the model (Brody et al., 2022; Gong & Cheng, 2019). As mentioned, edge attributes can hold multi-dimensional continuous information about the edge between two nodes, for example the synchrony per frequency band. Despite its potential, graphLambda has not yet been applied for the classification of psychiatric disorders, or EEG data in general, however, it has shown to have great potential for predicting the binding affinity of protein-ligand complexes (Mqawass & Popov, 2024).

Current literature describes the possibilities of EEG data and machine learning for the diagnosis of psychiatric disorders, in turn, providing insights into the underlying psychopathology. The novel deep learning architectures, such as graphLambda, have, however, not yet been extensively investigated for the classification of psychiatric disorders using EEG-derived features. Moreover, expanding on the knowledge gaps in the current literature, instead of multiclass classification algorithms, most research has investigated different binary classification algorithms for each psychiatric disorder separately, distinguishing between patients and HCs. Multiclass classification models could provide a more complete perspective on the predictive performance of certain ML approaches and EEG-derived features, by allowing the comparison of the underlying class predictions of a model. Additionally, multiclass classification models could provide an all-encompassing diagnostic methodology in the clinic for novel patients for which the clinician is not yet postulating any singular psychiatric disorder (e.g. ADHD or not). Lastly, current literature investigating ML approaches for the classification of psychiatric disorders typically derive their EEG features from a singular condition in which the EEG was measured, such as EO resting-state EEG data. EEG features derived from different

measurement conditions could result in different model performances, possibly providing valuable insight into the improvement of psychiatric disorder diagnosis, and underlying psychopathology.

To further investigate the potential utility of biomarkers in the current diagnosis of psychiatric disorders, and to obtain insights into the underlying psychopathology of these psychiatric disorders, the current article will examine the predictive performance of a multitude of ML approaches, trained with a multitude of different feature sets (Table 1). The classes to be classified include; ADHD, MDD, OCD, SMC, and HCs. The feature sets include; 1) statistical measurements, such as the standard deviation, of the TFR of an aggregated EEG signal of a group of electrodes for all frequency bands (delta, theta, alpha, beta, gamma), as well as 2) FC features, measured as the synchrony in phase between the aggregated EEG signals of electrode groups for all frequency bands. The conditions from which the two feature sets will be derived include EO resting-state EEG data, EC resting-state EEG data, and the ratio between the EO and EC resting-state EEG data. SVM, RF, and Gradient Boosting Classifier (GBC) models will be used as the standard ML models, and a standard GCN, a graphLambda without edge attributes, and a graphLambda with edge attributes will be used as the GCN models. The standard ML models will be trained with exclusively the statistical TFR features as a baseline, as well as with the statistical TFR features in conjunction with the FC features to investigate the additional predictive performance of the FC features (extended models). The GCN models will only be trained with the statistical TFR features in conjunction with the FC features, and not exclusively with the statistical TFR features. However, the standard GCN and the graphLambda without edge attributes are only able to utilize the FC features as binary encoded edges, based on an arbitrary FC threshold. Nonetheless, the graphLambda with edge attributes will be able to utilize the continuous FC features as input.

We expect that the inclusion of FC features in the extended ML models will result in better model performance, compared to the baseline models trained with exclusively the statistical TFR features. This is expected due to previous literature having shown that the FC in the brain is an important biomarker for ADHD, MDD, OCD, as well as disorders related to SMC, such as Alzheimer's disease. For the different EEG data conditions from which the two feature sets will be derived, we expect that, in general,

the EO- and EC-derived feature sets will lead to more accurate classification than the ratio-derived feature sets. We do expect, however, that the ratio-derived feature sets might show accurate classification for specific classes in which the difference between EO and EC resting-state activity is more apparent, such as in the power of the alpha frequency band in ADHD. For the GCN models, we expect that their ability to represent the spatial structure of the EEG data will result in better classification compared to all standard ML models. Moreover, we expect that the graphLambda models will outperform the standard GCN, due to the added GAT and GIN representations in the model architecture. Lastly, we expect that the addition of edge attributes in the graphLambda model will lead to better model performance, considering it includes the continuous synchrony values between brain regions.

Table 1 | Overview combinations of test conditions

Model type	Models	Conditions	Feature sets
Standard ML	SVM, RF, GBC (baseline)	EC, EO, ratio	Statistical TFR features
Standard ML	SVM, RF, GBC (extended)	EC, EO, ratio	Statistical TFR features + FC features
GCNs	Standard GCN, graphLambda without edge attributes, graphLambda with edge attributes	EC, EO, ratio	Statistical TFR features + FC features

Note. ML = Machine learning; GCN = Graph Convolutional Network; SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; EC = eyes-closed; EO = eyes-open; TFR = Time-Frequency Representation; FC = Functional connectivity

2 Methods

2.1 Database description

The data used for the analysis in the current article, is provided by the Two Decades-Brainclinic Research Archive for Insights in Neurophysiology (TDBRAIN) EEG database (van Dijk et al., 2022). The TDBRAIN database was presented in support of biomarker development and methodologies such as training machine learning models for data-driven neuroscientific research. The database contains EO and EC resting-state, raw EEG data of 1274 psychiatric patients and healthy controls (620 female, age 38.67 ± 19.21, range 5-89 years) collected between 2001 and 2021, complemented with relevant clinical and demographic data. The main psychiatric indications included are 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 treatment outcome data was blinded by the authors of the TDBRAIN database. The remaining sample (N = 1153) of the database was utilized for the current analysis. It is important to note that the TDBRAIN replication sample was not utilized in this study, considering the time-scope of the current project and the amount of correspondence needed for the evaluation by the Brainclinics foundation of the 27 different models. To still facilitate correct validation and replication practices, other actions were taken with the remaining sample (N = 1153), such as nested cross-validation and splitting off a hold-out set. These actions will be discussed in further detail in section *2.11 Model development*.

The EEG data were recorded during, one or multiple sessions, of a two-minute EO resting-state recording, in which participants focused on a red dot in a comfortable position, and a two-minute EC resting-state recording, in which participants remained in the same position with their eyes closed. Recordings consisted of 26 EEG electrodes, based on the 10-10 electrode international system, acquired at a sampling rate of 500 Hz with a 100 Hz low pass filter applied prior to digitization. Moreover, the recordings included electrodes recording vertical and horizontal eye movements, as well as electrocardiogram (ECG) and the electromyogram (EMG), for artifact detection. For further experimental conditions refer to Van Dijk et al. (2022).

A spreadsheet including the clinical information with diagnoses, demographic information, and additional practical information is provided alongside the EEG data. Psychiatric diagnoses in the TDBRAIN database were distinguished by either a

formal diagnosis, confirmed by a licensed clinician, or a referral-indication, meaning the participant was referred for the EEG assessment by a general practitioner or psychologist/psychiatrist. The formal diagnoses and indications include a variety of 39 single or comorbid psychiatric diagnoses, as well as unknown/missing entries and healthy controls (HCs). The value 'HEALTHY' was used to mark healthy participants, and unknown or missing entries were marked with 'UNKNOWN' or NaN. Participants with psychiatric disorders were marked with the respective disorder as the value, for example, participants with ADHD were marked with 'ADHD'. Finally, participants with comorbidity were marked with all respective disorders as the value, for example, participants with ADHD and OCD were marked with 'ADHD/OCD'

2.2 Software and packages

For the entire analytical pipeline, Python (version 3.11) was used. Raw EEG data were preprocessed using the script provided by the authors of the TDBRAIN database (van Dijk et al., 2022). Preprocessed EEG data were further processed to extract features with the open-source MNE-Python package (version 1.6) (Gramfort et al., 2013). Boruta feature selection was performed using the BorutaPy package (version 0.3) (Homola, 2015). Scikit-learn (version 1.4) was utilized to develop the standard ML models (Pedregosa et al., 2011). The neural networks were developed using Pytorch Lightning (version 2.2) and Optuna (version 3.6) (Akiba et al., 2019). Other Python packages used in the current analysis include, but are not limited to, NumPy (version 1.26), pandas (version 2.2), SciPy (version 1.13), and Matplotlib (version 3.8). Figures were edited using Inkscape editing software (version 1.2.2). Lastly, GitHub Copilot (version 1.7) was used for assistance in debugging and optimizing code.

2.3 Data analysis pipeline

A schematic overview of the analysis pipeline, from preprocessing to model evaluation, can be seen in Figure 1. In the next sections, all steps will be discussed in further detail. For exact code, refer to https://github.com/TSmolders/Internship_EEG/.

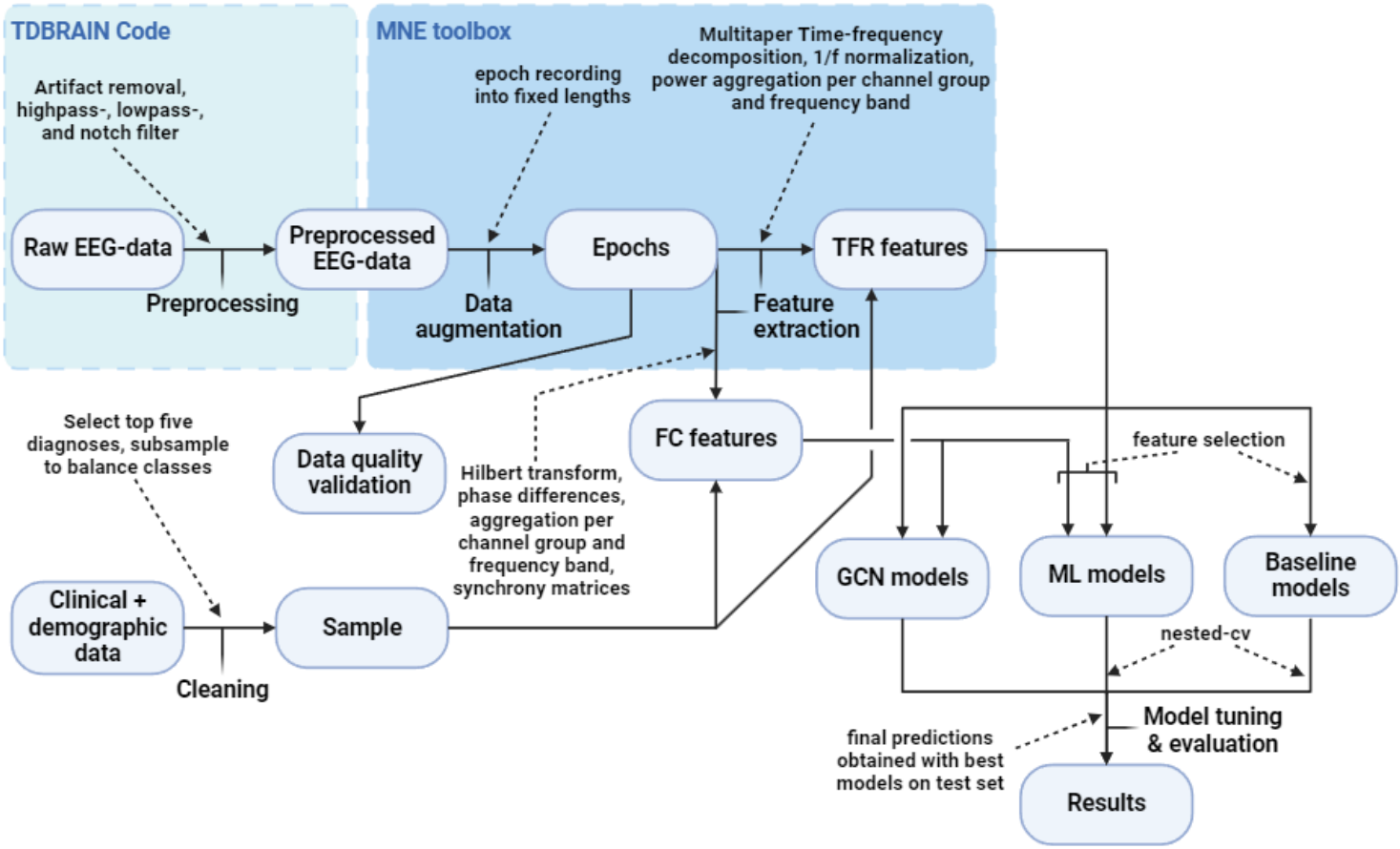


Figure 1 | Schematic overview analysis pipeline. Boxes represent objects within pipeline. General analytical approaches and direction are represented with solid arrows. Dashed arrows show specific steps within general approaches. Parts of the pipeline that were performed with the TDBRAIN code or the MNE toolbox have been encased in respectively labeled boxes. EEG = electroencephalography; TFR = time-frequency representation; FC = functional connectivity; GCN = graph convolutional network; ML = machine learning; cv = cross-validation.

2.4 Sampling

In order to develop reliable multiclass classification models, a large enough sample size is needed for each class. To assure enough training data, the formal diagnoses and indications were first aggregated to a single ‘diagnosis’ variable. This was done, by taking the formal diagnosis for the participant when known, and the indication when the formal diagnosis was unknown. To further assure enough instances were available per class, only the top five most frequent psychiatric disorders from the aggregated ‘diagnosis’ variable were taken: MDD (N = 323), ADHD (N = 176), SMC (N = 119), OCD (N = 49), HC (N = 47). Lastly, to minimize bias in model prediction, participants from the selected top five diagnoses were subsampled to balance the dataset. Subsampling was officially performed after feature extraction, to ensure that

subsampled entries had no to little missing variables, in turn ensuring that the dataset remained balanced throughout the pipeline. Due to the exclusion of two unusable artifact-ridden EEG recordings, the final sample consisted of 45 participants per psychiatric disorder after subsampling (Table 2).

Table 2 | Sample demographics

Diagnosis	N	Age (mean ± standard deviation)	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

2.5 Preprocessing

Raw EEG data was preprocessed using the provided pipeline from the authors of the TDBRAIN dataset (van Dijk et al., 2022). In this pipeline, firstly, the bipolar EOG was computed and removed from the EEG signal to eliminate eye-blink artifacts, according to the method published by Gratton et al. (1983). Next, the data was bandpass-filtered between 0.5 to 100 Hz, and notch-filtered at 50 Hz. After filtering, various artifacts in the signal were detected: EMG, sharp channel-jumps, kurtosis, extreme voltage swing, residual eyeblinks, electrode bridging, and extreme correlations. If more than 66% of the signal from an EEG channel contained artifacts, the channel's signal was repaired using a Euclidean distance weighted average of at least three neighboring channels. If the channel's signal could not be repaired, for example, due too many artifacts in neighboring channels, the entire EEG recording was marked as 'bad', and excluded from subsequent analyses.

2.6 Data augmentation

To artificially increase the size of our dataset, the preprocessed EEG recordings were segmented into 9.95-second epochs and used as independent observations with the same class label. Next to artificially increasing the sample size, segmenting

the recording was necessary considering the signal is non-stationary over longer periods and stationary over shorter periods. A stationary signal is important for obtaining interpretable results with wavelet convolution or multitaper analyses (M. X. Cohen, 2019). Furthermore, since there were no tasks during the resting-state EEG recording, the data were able to be divided into epochs of a fixed length without splitting task-evoked brain activity. A specific length of 9.95 seconds was used to obtain reliable statistical measurements of the TFR of a signal as features. This was necessary considering the wavelet and multitaper analyses that were performed in order to obtain the TFR, generally need at least three full cycles for reliable estimations. For example, for a frequency of 1 Hz, at least three seconds of data would have been needed. In turn, to ensure minimal overlap in statistical features from the TFR, a longer segment was used. Additionally, a specific length of 9.95 seconds was used to consistently yield twelve epochs for each EC and EO recording across all participants.

2.7 Data quality validation

Before proceeding with the analysis pipeline, a RF classifier was trained on Power Spectral Density (PSD) features extracted from the epoched EEG data to classify the data as either EO or EC recordings. Previous literature has shown that the power levels in EEG signals differ between EC and EO conditions, especially in the alpha band activity (Barry et al., 2007; Li et al., 2009). Thus, the quality of the preprocessed data could be validated by investigating the performance of a ML model to correctly classify the EO or EC condition of a recording. PSD estimates, expressed in squared micro-Volts per Hertz, were computed using Welch's method for all electrode groups (Figure 2) and for all frequency bands (Table 3). Data from singular electrodes that were not included in an electrode group were not used. Grouping the channels and frequencies was done to reduce the dimensionality of the data, and to resemble the feature extraction approach, discussed in section *2.8 Feature extraction*. The resulting PSD estimates for all electrode groups and all frequency bands were used to train a RF classifier with a K-fold cross-validation (CV) of 5 folds. The data from the K-fold CV folds were grouped among participants to ensure all epochs from the same participant retained to the same split. No optimization of the hyperparameters of the RF classifier was performed. Finally, the average accuracy of the CV results was evaluated, and the feature importances from

the RF were investigated to confirm the expected importance of the alpha power. The results were evaluated before starting with feature extraction, in order to ensure high quality features. The cross-validated RF was able to accurately classify 75% of EC or EO epochs (mean accuracy = 0.747 ± .016). In addition, the alpha band PSD features were the most important features in all CV folds, confirming the preprocessed data holds valuable information on brain activity and is of high quality.

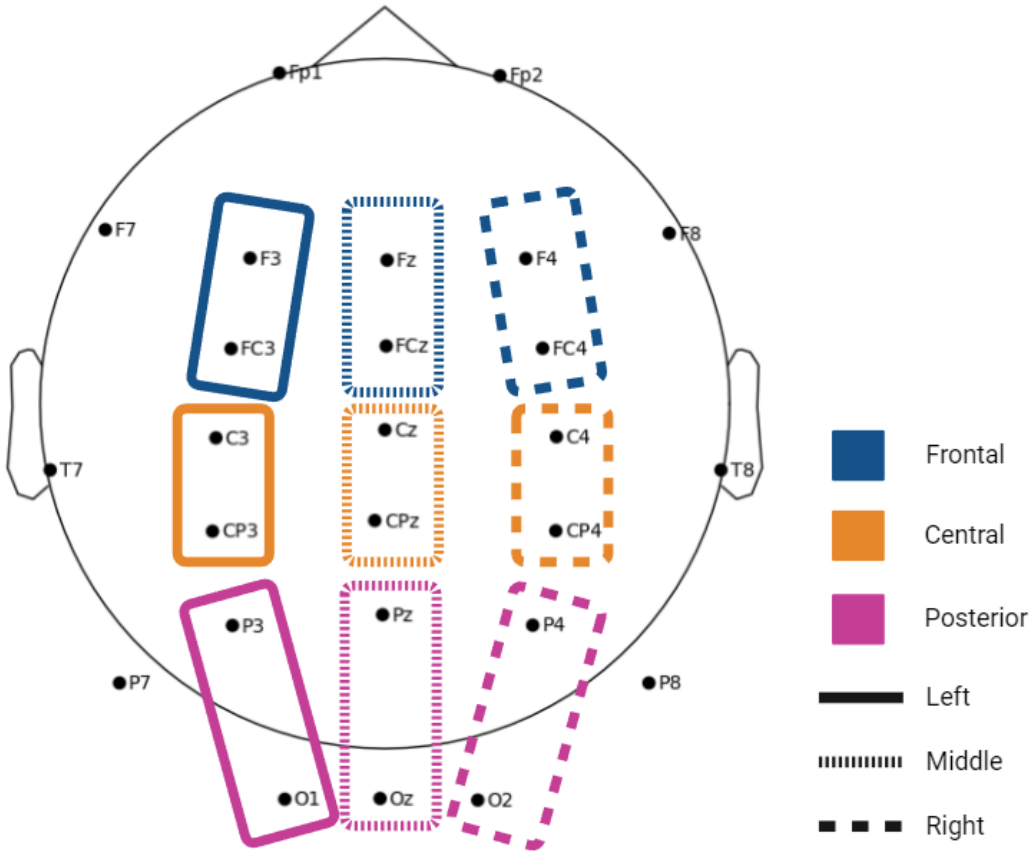


Figure 2 | Electrode montage and groups. Electrode locations are shown on a top-down view of a schematic head facing to the north. Electrodes belonging to the same electrode group are encased. All electrode groups are named according to their position on the head: Frontal, Central, or Posterior, in combination with Left, Middle, or Right, totaling nine different electrode groups. Electrodes not included in the nine electrode groups were not used for 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 is given, as well as the frequencies selected within each frequency band for the time-frequency representation.

2.8 Feature extraction

The current analysis included multiple feature sets. Firstly, we extracted features separately for the EC resting-state EEG data, and the EO resting-state EEG data. Additionally, we created a third condition by dividing the computed TFR of the EC condition by the computed TFR of the EO condition, before extracting statistical features from this 'ratio' signal. For the FC features, the 'ratio' condition was computed by dividing the actual EC FC features by the EO FC features, as opposed to dividing the signals. This approach aimed to examine the predictive power of features representing the differences between EC and EO brain activity for diagnosing psychiatric disorders. For each of these three conditions, two different feature sets were extracted: 1) statistical features (mean, standard deviation, median, skewness, kurtosis) from a TFR of the epoch, and 2) FC features of the synchrony between electrode groups. Similar to the PSD estimates in the data quality validation, all features were aggregated over electrode groups (Figure 2) and frequency bands (Table 3). Electrodes not belonging to an electrode group were not included. The next subsections will detail the feature extraction methods for the respective feature sets. To investigate the difference in classification performance for psychiatric disorders with the addition of FC features, the FC feature set was combined with the statistical TFR feature set. This resulted in two different feature sets for each of the three conditions (EC, EO, ratio): 1) statistical TFR features, and 2) statistical TFR features + FC features.

2.8.1 Statistical time-frequency representation features

The statistical TFR features were extracted by first computing the TFR of an epoch, including all electrodes within electrode groups (Figure 2), and all selected frequencies within all bands (Table 3). The TFR was computed using the multitaper method, which is useful for non-phase-locked and non-time-locked data, by smoothing over the finer details or noise while highlighting the prominent features within a signal (Babadi & Brown, 2014). Smoothing the noise and highlighting broader features within a signal is especially valuable in our current analysis, due to the uneventful nature of the resting-state EEG data. The multitaper method utilizes a number of discrete prolate Slepian sequences (DPSS), which are orthogonal tapers, to taper a signal a multitude of ways (Babadi & Brown, 2014). A fast Fourier transform (FTT) is then applied to these tapered signals, and the results are averaged together to obtain a single power spectrum. This process is repeated over small frequency-specific temporal windows over the entire epoch to obtain a time-frequency representation. After obtaining the epoched TFR signal with the multitaper method, it was normalized to a relative TFR by dividing each sample through the sum of the average power over time per frequency, calculated from the same epoch. Thus, the average power across frequencies and timepoints of the normalized TFR summed up to 1. For the ratio condition, the normalized TFR epoch of the EC recording was divided by the normalized TFR epoch of the EO recording. Next, the resulting TFRs for all conditions, were aggregated over electrode groups and frequency bands by averaging the signal over the respective electrodes and frequencies within an electrode group and frequency band. This aggregated TFR was then used to compute the following statistical features of the power over time for all electrode groups and frequency bands: mean, standard deviation, median, skewness, kurtosis. This resulted in a total of 225 (9 electrode groups * 5 frequency bands * 5 statistical features) statistical TFR features per epoch for each condition.

2.8.2 Functional connectivity features

The consistency of the phase relation between two signals can provide an estimation of synchrony, and in turn, FC. Here, synchrony estimations were extracted between each electrode group (Figure 2) per frequency band (Table 3), resulting in a total of 405 (5 frequency bands * 81 possible connections) FC features. Of these 405 FC features, 45 features (5 frequency bands * 9 electrode groups) represented self-connected synchrony of electrode groups, encoded as a value of 1. Self-connected

values were specifically included for the GCNs to allow the input to a node in a hidden layer to depend on itself in addition to its neighbors. To compute the FC features, each signal was first band-pass filtered with the upper and lower frequencies of each respective frequency band. A Hilbert transformation (Hilbert, 1912) was then performed to obtain the phase of the signal, represented as an angle, per frequency band per electrode. Subsequently, the difference in the phase between two signals, computed by calculating the difference between two angles, was calculated to obtain a synchrony estimation for all possible bidirectional electrode connections per frequency band. Next, these synchrony estimations per channel per frequency band were aggregated by electrode group, by averaging over the respective electrodes within an electrode group (Figure 2). Before moving on to feature selection, the FC features for the standard ML models were filtered to include only one direction of the bidirectional synchrony estimations and exclude the self-connected values of the electrode groups, reducing the number of FC features to a total of 180 (5 frequency bands * 36 possible connections). This was exclusively done for the FC features for the standard ML models, since the full bidirectional data is vital for the graphical input of the GCN models, as opposed to the standard ML models where the input is flattened.

2.9 Feature selection

Through feature extraction, two different feature sets were obtained (1. statistical TFR features, 2. statistical TFR features + FC features) for three different conditions (EC, EO, ratio). However, the dimensionality of the data posed a risk of overfitting for the standard ML models (SVM, RF, GBC), which, as opposed to GCNs, cannot inherently reduce the dimensionality. Therefore, Boruta feature selection was performed before ML model development. Boruta is an ‘all-relevant’ feature selection wrapper algorithm that, in the current case, wraps a RF (Kursa & Rudnicki, 2010). ‘All-relevant’ feature selection identifies all relevant features, instead of only the minimal subset of features which are performant in a model, as done in ‘minimal-optimal’ feature selection. Selecting all relevant features is especially important when one is interested in understanding mechanisms related to the subject of interest. For each feature set from each condition, Boruta feature selection was performed separately. First, the ‘maximum depth’ and the ‘number of estimators’ hyperparameters for each Boruta-wrapped RF were tuned by finding at which

hyperparameter values the RF classifier achieves convergence at minimal value of the Out-of-bag (OOB) error, as described by Kursa & Rudnicki (2010). The tuned hyperparameters were then used for feature selection of their respective feature set and condition. The number of selected features per feature set and condition is shown in Table 4. For a more detailed view of the selected features, see Appendix A.

Table 4 | Feature selection overview (selected features / total features)

Feature set	Condition		
	EC	EO	Ratio
Statistical TFR features	147 / 225	135 / 225	66 / 225
Statistical TFR features + FC features	319 / 405	301 / 405	182 / 405

Note. TFR = Time-Frequency Representation; FC = functional connectivity; EC = eyes-closed; EO = eyes-open

2.10 Cluster analysis

To attempt to find class clusters within feature sets, cluster analyses were performed for each Boruta-selected feature set from each condition. If clear separation between class-specific clusters was found in the high-dimensional feature sets, it would suggest that the ML models could effectively learn to distinguish between the classes. The cluster analyses were performed using t-distributed Stochastic Neighbor Embedding (t-SNE). t-SNE is an unsupervised non-linear dimensionality reduction technique, producing significantly better visualizations than other techniques, such as Stochastic Neighbor Embedding, by reducing the tendency to crowd points together in the center of the map (Cieslak et al., 2020; Maaten & Hinton, 2008; Taskesen & Reinders, 2016). 2D and 3D visualizations were created using t-SNE, with a few different values for perplexity, which can be interpreted as a smooth measure of the effective number of neighbors (Maaten & Hinton, 2008). Next, the data points in the visualizations were colored according to their respective classes to investigate the class clustering within each feature set. However, no clear clustering of the psychiatric disorders was visible within the feature sets in any of the visualizations (Appendix B), indicating potential difficulties for models in distinguishing the classes based on the provided features.

2.11 Model development

Several types of models were built for the multiclass classification problem. For each condition (EC, EO, ratio), models can be divided into three categories: 1) standard ML models trained with statistical TFR features (baseline), 2) standard ML models trained with statistical TFR features and FC features (extended), 3) GCNs trained with statistical TFR features + FC features. The next subsections will describe the development of the models in these categories.

2.11.1 Standard ML models

The standard ML models utilized in the current study were SVM, RF, and GBC. These models were selected to include conceptually different algorithmic approaches to prediction, namely margin learning and ensemble learning, thus including a broader perspective on the strengths and weaknesses of ML algorithms in our analysis. Additionally, these algorithms were specifically selected due to their performance in EEG-based prediction problems in previous literature (Mari et al., 2022; Saeidi et al., 2021). Each ML model was trained with the two types of Boruta-selected feature sets for each of the three conditions (EC, EO, ratio), resulting in a total of 18 separate models. Before training, the features were standardized by subtracting the mean of the feature in the training set and dividing it by the standard deviation of the feature in the training set. The mean and standard deviation of the training set, instead of the entire sample, was used to prevent the distribution of the test set leaking into the model. After standardization, twelve missing entries were mean imputed for the EC and ratio feature sets due to a missing EC recording from a single participant.

2.11.1.a Hyperparameter tuning & model evaluation

To minimize the risk of bias in the model performance due to hyperparameter tuning, nested cross-validation (CV) was performed for all ML models. The nested CV consisted out of three inner splits and five outer splits (Figure 3). Stratified and grouped K-fold data splitting was applied in both the inner and outer loops of nested CV to ensure balanced classes among splits and to maintain epochs from the same participant within the same split. In the inner loop, hyperparameter tuning was performed by optimizing cross-validated F1-scores. A grid-search was used for the SVM models, while a randomized-search was used for the RF and GBC models due to their longer computation time, which was less of a limiting factor for the SVMs. See Table 5 for the hyperparameters and the search ranges. A mean F1-score

across folds was reported, resulting from the five cross-validated F1-scores for each outer split, and subsequently used for model evaluation. Additionally, for each standard ML model, a one-tailed paired t-test was performed between the F1-score distribution of the five cross-validated F1-scores of the baseline model compared to the F1-score distribution of the respective extended model. The alternative hypothesis posited that the extended model obtained higher F1-scores than the baseline model. Lastly, for all ML models, the tuned model with the highest F1-score from the five tuned models in the nested CV was refitted on the entire dataset to obtain predictions to compute a confusion matrix, as well as the precision, recall, and F1-score per class. Predictions were obtained using a non-nested K-fold CV approach, in which the dataset is split into test and train sets, with each sample belonging to exactly one test set, and its prediction computed with the estimator fitted on the corresponding train set. The same outer CV loop (n = 5) for the nested CV was used. The underlying class predictions were obtained to investigate the underlying capabilities of the model to predict specific psychiatric disorders.

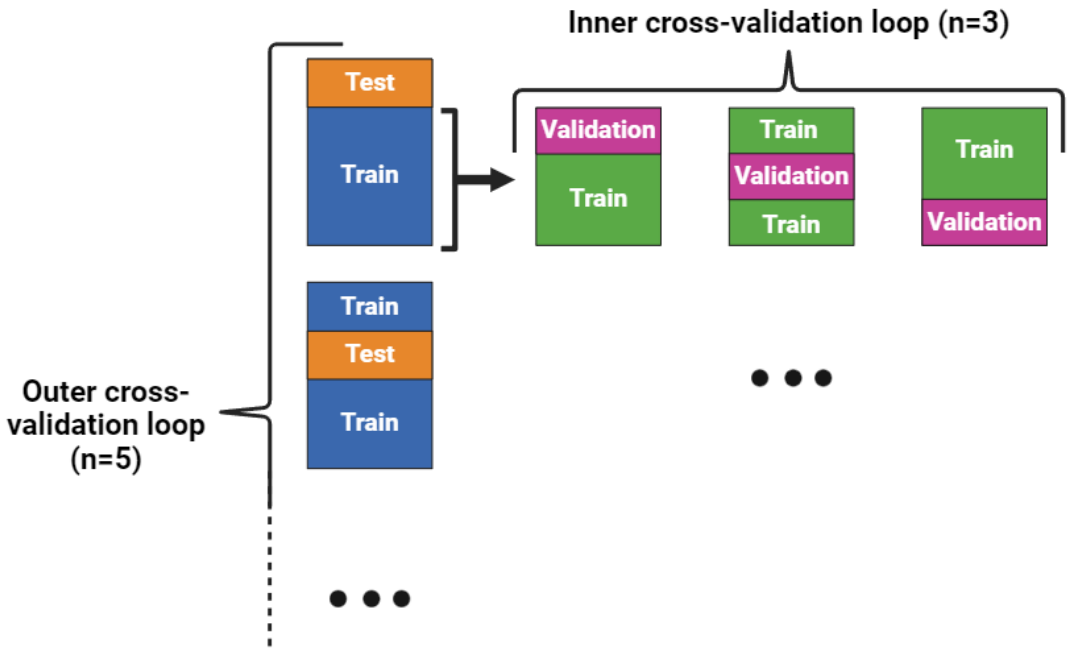


Figure 3 | Schematic overview of nested cross-validation. In the outer cross-validation loop, the data set is initially split into five different train (blue) and test (orange) sets. Subsequently, the five different train sets are split into three different train (green) and validation (pink) sets in the inner cross-validation loop. Three dots and the dashed line denote continuation.

Table 5 | The hyperparameters of the standard ML models and the search ranges during optimization.

Algorithm	Hyperparameter search ranges		
SVM	*C:*	*gamma:*	*kernel:*
	[0.01, 0.1, 1, 10, 100]	[0.1, 1, 10, 100]	['rbf']
RF	*n estimators:*	*max. features:*	*max. depth:*
	[200, 400, 600 … 2000]	['log2', 'sqrt']	[10, 20, 30 … 110, None]
	min. samples per split:	*min. samples per leaf:*	*bootstrap:*
	[2, 5, 10]	[1, 2, 4]	['True', 'False']
GBC	*n estimators:*	*loss:*	*learning rate:*
	[100, 200, 300 … 1000]	['log_loss']	[loguniform(1e-3, 1e-1, 1000)]*
	subsample:	*criterion:*	*min. samples per split:*
	[0.5, 0.6, 0.7, … 1]	['friedman_mse']	[2, 5, 10]

Note. The hyperparameters of the SVM were tuned using a grid-search approach, while the hyperparameters of the RF and GBC 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.

2.11.2 GCN models

Three different GCN architectures were utilized for the current analysis: 1) a standard GCN, 2) graphLambda without edge attributes, and 3) graphLambda with edge attributes. These models were selected to obtain an insight into the performance of GCNs ranging in complexity. Due to the graphical nature of the GCNs, the GCN models were only trained with the 'statistical TFR features + FC features' feature set, for each of the three conditions (EC, EO, ratio), totaling 9 different models. Considering the convolutional layers in the GCNs providing a built-in dimensionality reduction approach, the GCN models were trained with the feature

sets before Boruta feature selection. The input for all GCN models consisted of node features, which were the statistical TFR features for each channel group, and the edge indices, which encode the connections between nodes (Figure 4). The edge indices in the standard GCN and graphLambda without edge attributes, were excluded or included based on a minimum synchrony threshold, which was a hyperparameter during tuning. The graphLambda with edge attributes allowed for the inclusion of the synchrony features as continuous input, in turn, not needing to exclude edge indices based on a threshold. In the next subsections, the architectures of these GCNs are explained in further detail.

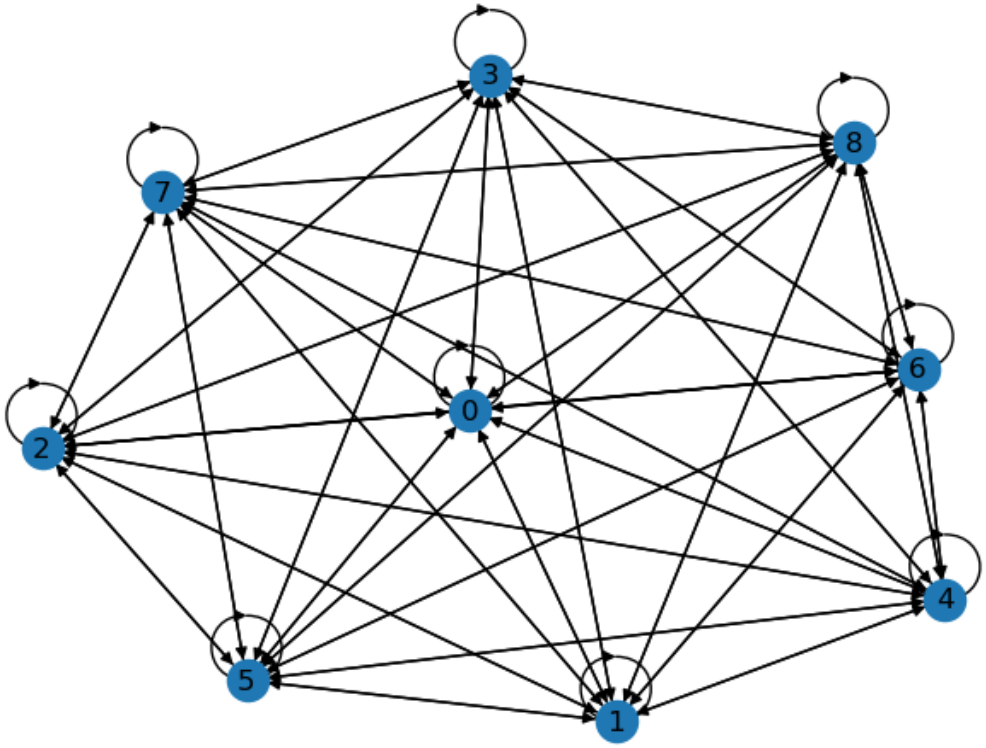


Figure 4 | Example of graphical input for graph convolutional networks. Electrode groups are represented as nodes (blue circles) and connections between the electrode groups are represented as edges (arrows), the direction of the connections is visualized by arrow heads. Each node contains node features, and each edge is either included or excluded by the edge index. When included, edges can contain continuous edge attributes. In this example, all possible edges are included.

2.11.2.a Standard GCN

The standard GCN architecture consisted out of two convolutional layers, and three fully-connected (FC) layers, in that order. The two convolutional layers and the first FC layer were each followed by a batch normalization (BN) layer and a dropout (DO) layer. Relu activation functions were used, as well as a log softmax activation function for the final layer. After each training step, loss was computed using the negative log likelihood (NLL) loss function.

2.11.2.b graphLambda

Both graphLambda architectures consisted of a GCN representation, a GAT representation, and a GIN representation, with each representation including BN layers. These representations are concatenated and put through FC layers, including DO layers. The amount of FC layers, and therefore DO layers, was a hyperparameter during tuning. All GCN-, GAT-, GIN-, and FC layers, except the last FC layer, used Relu activation functions. The final FC layer used a log softmax activation function. Loss was computed using the NLL loss function. The only difference in model architecture between graphLambda without edge attributes and with edge attributes, is that the GAT representation of the graphLambda with edge attributes allows for the input of edge attributes.

2.11.2.c Hyperparameter tuning & model evaluation

Due to computational restraints, it was not possible to perform nested CV for the development of the GCNs. In turn, the obtained model performances are at risk of being slightly biased. Nevertheless, hyperparameter tuning was performed using a single split taken from one of the outer and inner splits of the nested CV during ML model development, in order to facilitate an as fair as possible comparison in performance. The outer split was used to obtain the test set, and the inner split was used to obtain the train and validation set. The train set was used for training the models, the validation set for evaluating the models during tuning, and the test set for final model evaluation after tuning. For all GCNs, 100 trials of a randomized-search approach were performed, optimizing the F1-score by finding the optimal learning rate, optimizer, dropout percentage, and batch size (Table 6). In addition, the number of FC layers was included as a hyperparameter during tuning of the graphLambda models, as well as the synchrony threshold for the inclusion of edges in the input of the standard GCN and graphLambda without edge attributes. The number of training epochs was varied depending on the batch size for each trial to

obtain a similar amount of training steps for each trial. To minimize computational time, trials of the graphLambda models were stopped early if the intermediate F1-score was still lower than 0.1 after 100 training steps. For the standard GCN models, trials were stopped early if the intermediate F1-score was lower than the median final F1-score of the previous trials. Intermediate F1-scores and loss on the validation set and the training set were retrieved to investigate model quality (Appendix C). Model checkpoints of the best tuned model for each GCN type were retrieved to obtain final evaluation on the test set, including a confusion matrix and the precision, recall, and F1-score for each individual class, which was reported and subsequently used for model evaluation.

Table 6 | Hyperparameters of the GCN models and the search ranges during optimization.

Hyperparameter	GCN model		
	Standard GCN	GL wo edge attributes	GL with edge attributes
Learning rate[a]	[Loguniform(1e-5, 1e-1)]	[Loguniform(1e-5, 1e-1)]	[Loguniform(1e-5, 1e-1)]
Optimizer	['Adam', 'RMSprop', 'SGD']	['Adam', 'RMSprop', 'SGD']	['Adam', 'RMSprop', 'SGD']
Batch size	[170, 340, 510, 680, 850]	[170, 340, 510, 680, 850]	[170, 340, 510, 680, 850]
Dropout chance[b]	[0 … 0.8]	[0 … 0.8]	[0 … 0.8]
Edge threshold[b]	[0.5 … 0.9]	[0.5 … 0.9]	-
FC layers	-	[1, 2, 3 … 8]	[1, 2, 3 … 8]

Note. All hyperparameters were optimized using a randomized-search approach.

[a]The search range is a loguniform distribution within 1e-5 to 1e-1.

[b]Only the minimum and maximum values of the search range for this hyperparameter were assigned, any float value within this range could be sampled.

3 Results

In the following section, the results regarding the investigation of the predictive power for multiclass psychiatric disorder classification, for different feature sets and models, will be described. First, an overview of the performance of all models will be investigated, after which the performance of individual model types will be described in more detail.

3.1 Overview model performance

Table 7 | Model performance in F1-score

Models	EC		EO		Ratio	
	Baseline	Extended	Baseline	Extended	Baseline	Extended
SVM	.267 ± .032	.091 ± .019	.326 ± .048	.141 ± .034	.234 ± .021	.181 ± .025
RF	.361 ± .016	.411 ± .027**	.365 ± .074	.394 ± .055	.277 ± .051	.301 ± .054
GBC	.370 ± .027	**.440 ± .052***	.383 ± .060	.402 ± .052*	.266 ± .044	.308 ± .044*
Standard GCN	-	.303	-	.375	-	.249
GL without edge attributes	-	.374	-	.366	-	.256
GL with edge attributes	-	.362	-	.352	-	.261

Note. The SVM, RF, GBC F1-scores are presented as mean ± standard deviation from the nested cross-validation. Baseline models were trained with exclusively statistical time-frequency representation features, while the extended models were trained with statistical time-frequency features in conjunction with functional connectivity features. The F1-scores for the standard GCN and the GL with and without edge attributes are from the singular best model during tuning. The highest obtained F1-score distribution is indicated in bold. One-tailed paired t-tests were performed to test if the F1-scores of the extended model were greater than the F1-scores of the respective baseline model. EC = eyes-closed; EO = eyes-open; SVM = Support Vector Machine; RF = Random Forest; GCN = Graph Convolutional Network; GL = graphLambda.

*p-value < 0.05, **p-value <0.01

The performance of all models, developed with the different feature sets and conditions, was characterized by the average F1-scores of the CV folds from the ML models, as well as the best F1-scores of the GCNs obtained during tuning. Table 7 shows that, except for the SVM and the RF trained with EO- and ratio-derived features, the standard ML models trained with the FC features in conjunction with the statistical TFR features (extended) obtain a significantly higher F1-score compared to the models trained with exclusively the statistical TFR features (baseline). In contrast, the best F1-scores of the GCN models are within the F1-score distributions of the baseline models. Overall, the F1-scores obtained with the extended standard ML models are higher than the F1-scores obtained with the more complex GCN models. Moreover, the complexity of the graphLambda models, including the use of edge attributes, does not seem to result in a higher F1-score when comparing to the standard GCN. Additionally, feature sets extracted from EC and EO recordings generally yielded higher F1-scores than those derived from the ratio between EC and EO recordings. This holds true for all models, except the extended SVM models. Notably, the GBC model trained with EC-derived statistical TFR features and FC features attained the highest F1-score. However, in general there was no consistent superiority of either EC- or EO-derived feature sets when comparing the distributions of the F1-scores.

3.2 Standard ML models

The recall, precision, and F1-score for each individual class, as well as a confusion matrix, were computed with the best tuned model during model development with nested CV. This was done for each feature set and each condition for all standard ML models to investigate the underlying characteristics in predictions of the model. Results from models trained with EC- and EO-derived feature sets generally indicate the highest accuracy in predicting ADHD, SMC, and HC classes. Conversely, models trained with ratio-derived feature sets demonstrate limited ability to correctly classify classes, except for ADHD.

3.2.1 EC-derived feature sets

Table 8 | Class-specific F1-scores of ML models trained with EC-derived feature sets

Class label	SVM		RF		GBC	
	Baseline	Extended	Baseline	Extended	Baseline	Extended
ADHD	0.35	0.04	0.46	0.49	0.46	0.55
HC	0.29	0.20	0.43	0.46	0.42	0.43
MDD	0.19	0.03	0.27	0.34	0.28	0.36
OCD	0.15	0.01	0.31	0.33	0.33	0.38
SMC	0.38	0.33	0.40	0.44	0.45	0.55

Note. Baseline models were trained with exclusively statistical time-frequency representation features, while the extended models were trained with statistical time-frequency features in conjunction with functional connectivity features. The class-specific F1-scores are obtained from the best tuned model during nested cross-validation. EC = eyes-closed; SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

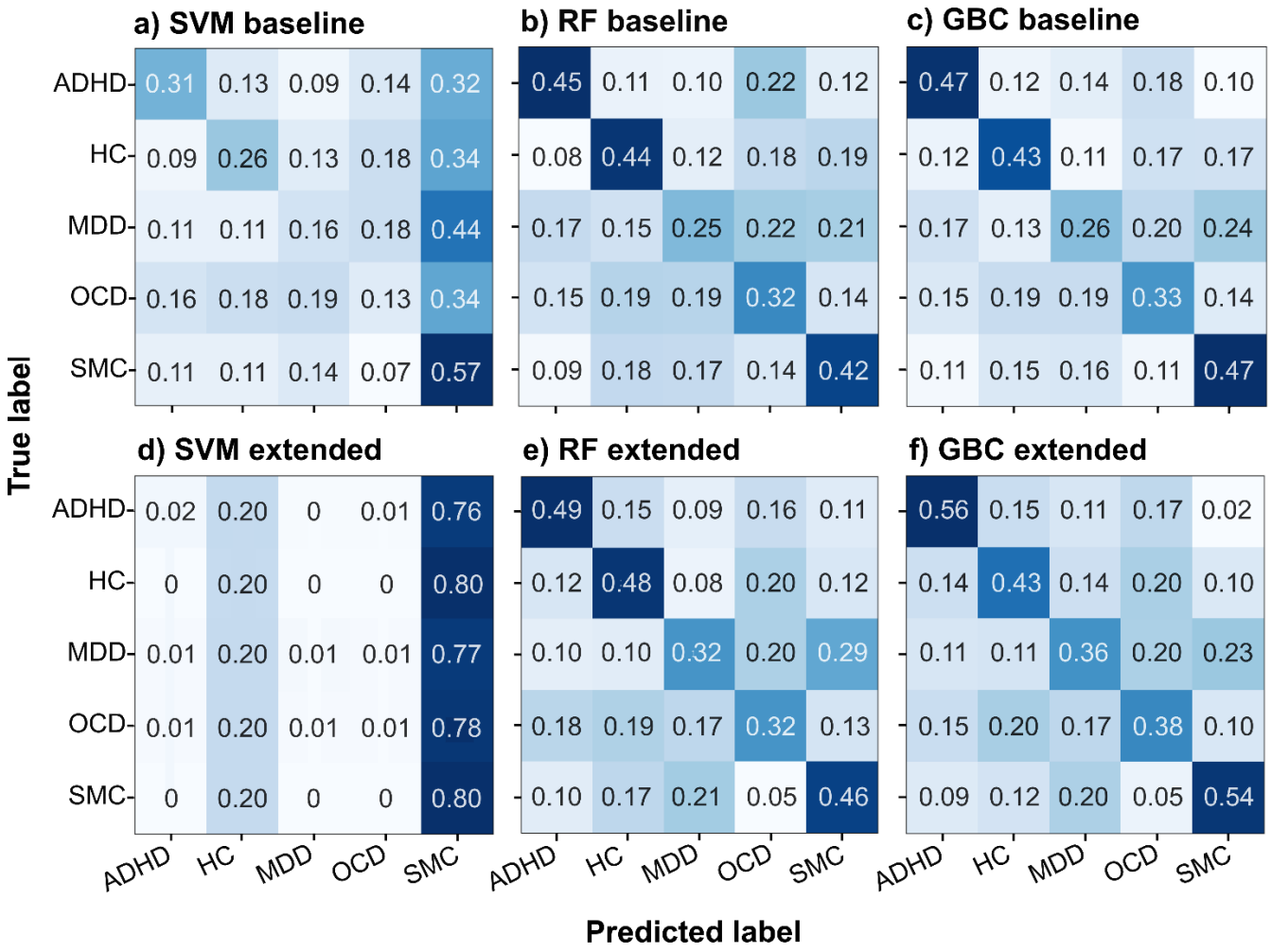


Figure 5 | Confusion matrices from ML models trained with EC-derived feature sets. Confusion matrices obtained from a single fold of the nested cross-validation are shown for the baseline models (a – c) and the extended models (d – f), which were trained with the addition of functional connectivity features. 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. For the absolute values see Appendix D. SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

The RFs and GBCs of the standard ML models trained with the EC-derived feature sets, in general, show that ADHD is most accurately distinguished from the other classes, followed by HC and SMC (Table 8 & Figure 5). In contrast, MDD and OCD show the least accurate classification. Moreover, the results show that the F1-scores of all classes increase with the addition of FC features in the RF and GBC models (Table 8). The SVM models seem to (mis)classify most epochs as SMC, with some

distinction between ADHD, HC, and SMC, which becomes worse with the addition of the FC features. Disregarding the SVM models, ADHD and HC are most often misclassified with OCD, and OCD misclassified with HC and MDD (Figure 5). Notably, MDD is overall most often misclassified with SMC, and SMC with MDD, as well as HC in the baseline models.

3.2.2 EO-derived feature sets

Table 9 | Class-specific F1-scores of ML models trained with EO-derived feature sets

Class label	SVM		RF		GBC	
	Baseline	Extended	Baseline	Extended	Baseline	Extended
ADHD	0.44	0.24	0.48	0.50	0.49	0.47
HC	0.31	0.03	0.37	0.37	0.39	0.35
MDD	0.18	0.02	0.21	0.21	0.28	0.26
OCD	0.25	0.13	0.33	0.36	0.35	0.42
SMC	0.46	0.37	0.42	0.51	0.44	0.54

Note. Baseline models were trained with exclusively statistical time-frequency representation features, while the extended models were trained with statistical time-frequency features in conjunction with functional connectivity features. The class-specific F1-scores are obtained from the best tuned model during nested cross-validation. EO = eyes-open; SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

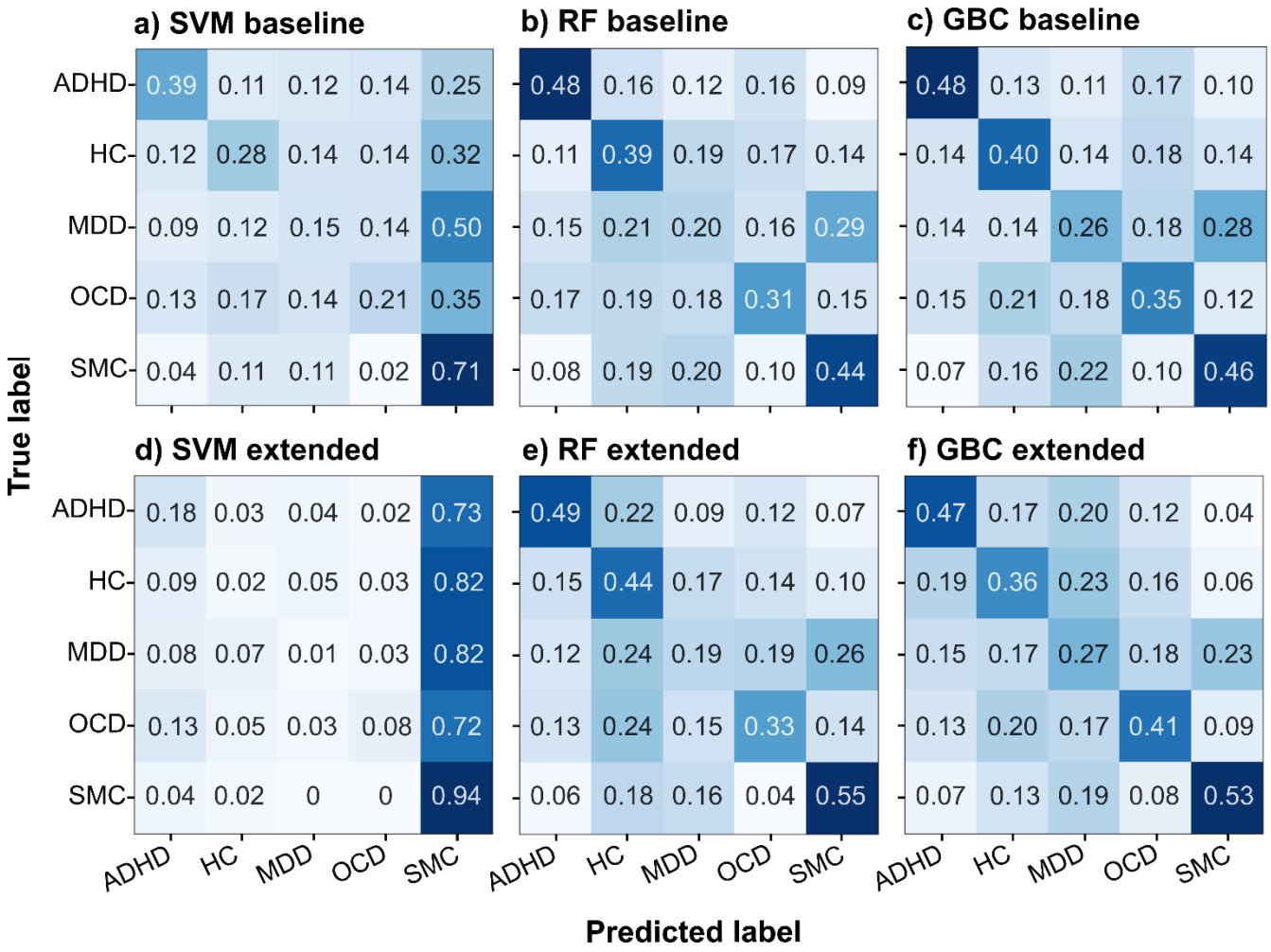


Figure 6 | Confusion matrices from ML models trained with EO-derived feature sets. Confusion matrices obtained from a single fold of the nested cross-validation are shown for the baseline models (a – c) and the extended models (d – f), which were trained with the addition of functional connectivity features. 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. For the absolute values see Appendix D. SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

The RFs and GBCs trained with the EO-derived feature sets, in general, most accurately classify ADHD in the baseline models, and SMC in the extended models trained with the addition of the FC features (Table 9 & Figure 6). However, the addition of the FC features does not seem to yield a large improvement of the F1-scores of HC, and MDD. Notably, the F1-score of OCD does seem to improve for the extended GBC model, and to a lesser extent as well for the RF extended model,

compared to the baseline GBC and RF models (Table 9). Again, the SVM models seem to (mis)classify most cases as SMC, with some distinction between ADHD, HC, and SMC, which worsens with the addition of the FC features (Figure 6). When isolating the RF and GBC models, the confusion matrices show that MDD is most often misclassified with SMC and vice versa, as well as with HC in the RF models.

3.2.3 Ratio-derived feature sets

Table 10 | Class-specific F1-scores of ML models trained with ratio-derived feature sets

Class label	SVM		RF		GBC	
	Baseline	Extended	Baseline	Extended	Baseline	Extended
ADHD	0.36	0.09	0.43	0.41	0.39	0.39
HC	0.24	0.28	0.29	0.33	0.28	0.28
MDD	0.19	0.14	0.17	0.24	0.15	0.26
OCD	0.21	0.23	0.25	0.27	0.29	0.30
SMC	0.25	0.26	0.26	0.33	0.25	0.33

Note. Baseline models were trained with exclusively statistical time-frequency representation features, while the extended models were trained with statistical time-frequency features in conjunction with functional connectivity features. The class-specific F1-scores are obtained from the best tuned model during nested cross-validation. EC = eyes-closed; SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

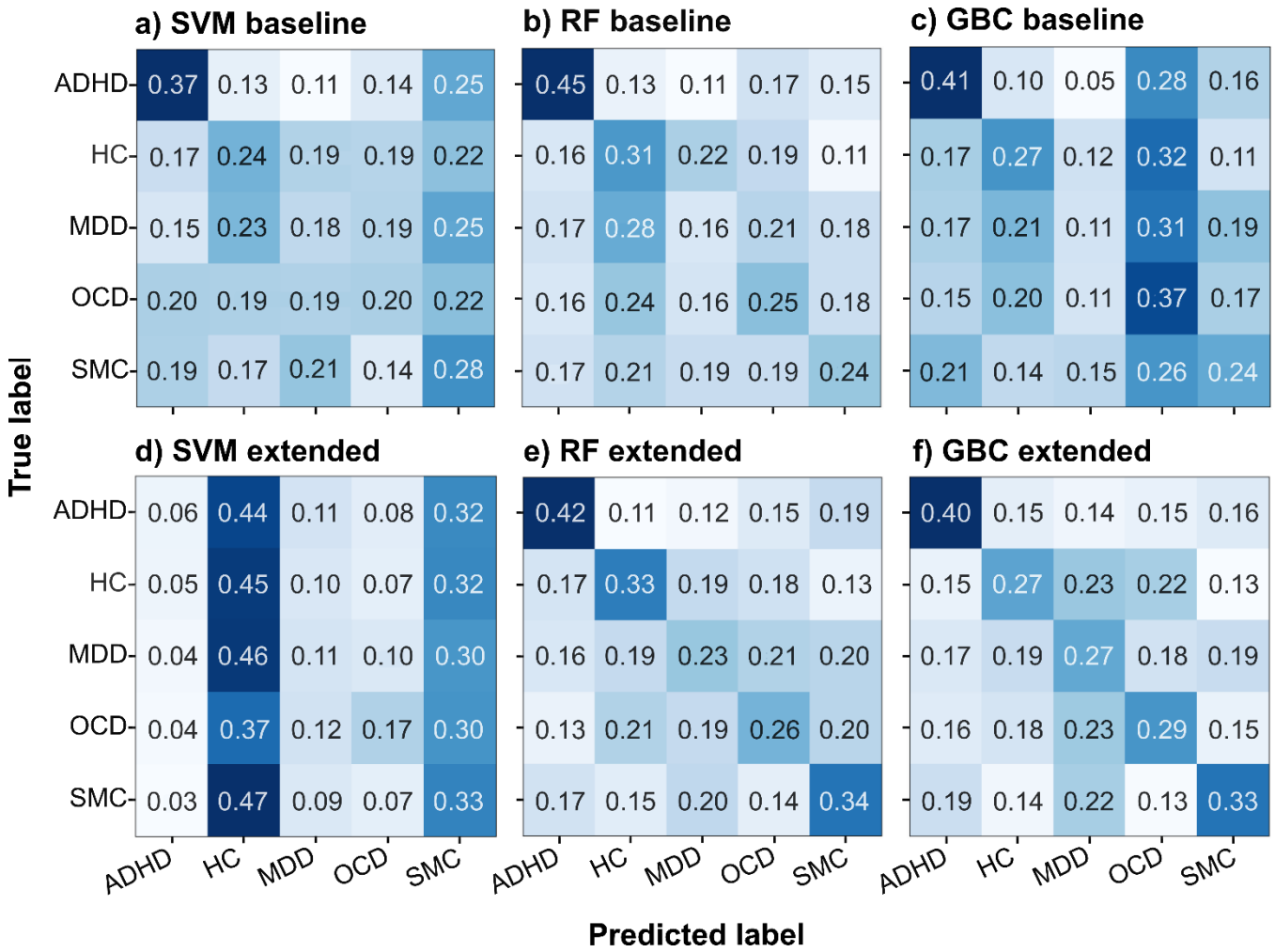


Figure 7 | Confusion matrices from ML models trained with ratio-derived feature sets. Confusion matrices obtained from a single fold of the nested cross-validation are shown for the baseline models (a – c) and the extended models (d – f), which were trained with the addition of functional connectivity features. 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. For the absolute values see Appendix D. SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

All models, except for the extended SVM model, trained with the ratio-derived feature sets, seem to most accurately classify ADHD (Table 10 & Figure 7). However, the results show that the baseline models are not able to accurately distinguish between the other classes. Nevertheless, the addition of the FC features to the RF and GBC models does show to improve the distinction between the classes, as represented by an increase in the class-specific F1-scores. The addition of the FC

features to the SVM model generally results in worse distinction between the classes.

3.3 GCN models

The recall, precision, and F1-score for each individual class, as well as a confusion matrix, were computed by obtaining predictions on the held-out test set with the best model during tuning of each GCN type. This was done for each feature set and each condition, to investigate the underlying characteristics in predictions of the model.

Table 11 | Class-specific F1-scores of GCN models

Class label	Standard GCN			GL without edge attrs			GL with edge attrs		
	EC	EO	Ratio	EC	EO	Ratio	EC	EO	Ratio
ADHD	0.46	0.49	0.42	0.39	0.47	0.39	0.46	0.43	0.35
HC	0.42	0.46	0.22	0.47	0.38	0.19	0.47	0.37	0.28
MDD	0.27	0.23	0.10	0.32	0.33	0.19	0.23	0.28	0.14
OCD	0.21	0.33	0.26	0.33	0.38	0.28	0.34	0.35	0.24
SMC	0.15	0.36	0.24	0.35	0.27	0.23	0.31	0.33	0.29

Note. The class-specific F1-scores are obtained from the predictions on the test set from the best model during tuning, for each GCN model. EC = eyes-closed; EC = eyes-closed; EO = eyes-open; GCN = Graph Convolutional Network; GL = graphLambda; attrs = attributes; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

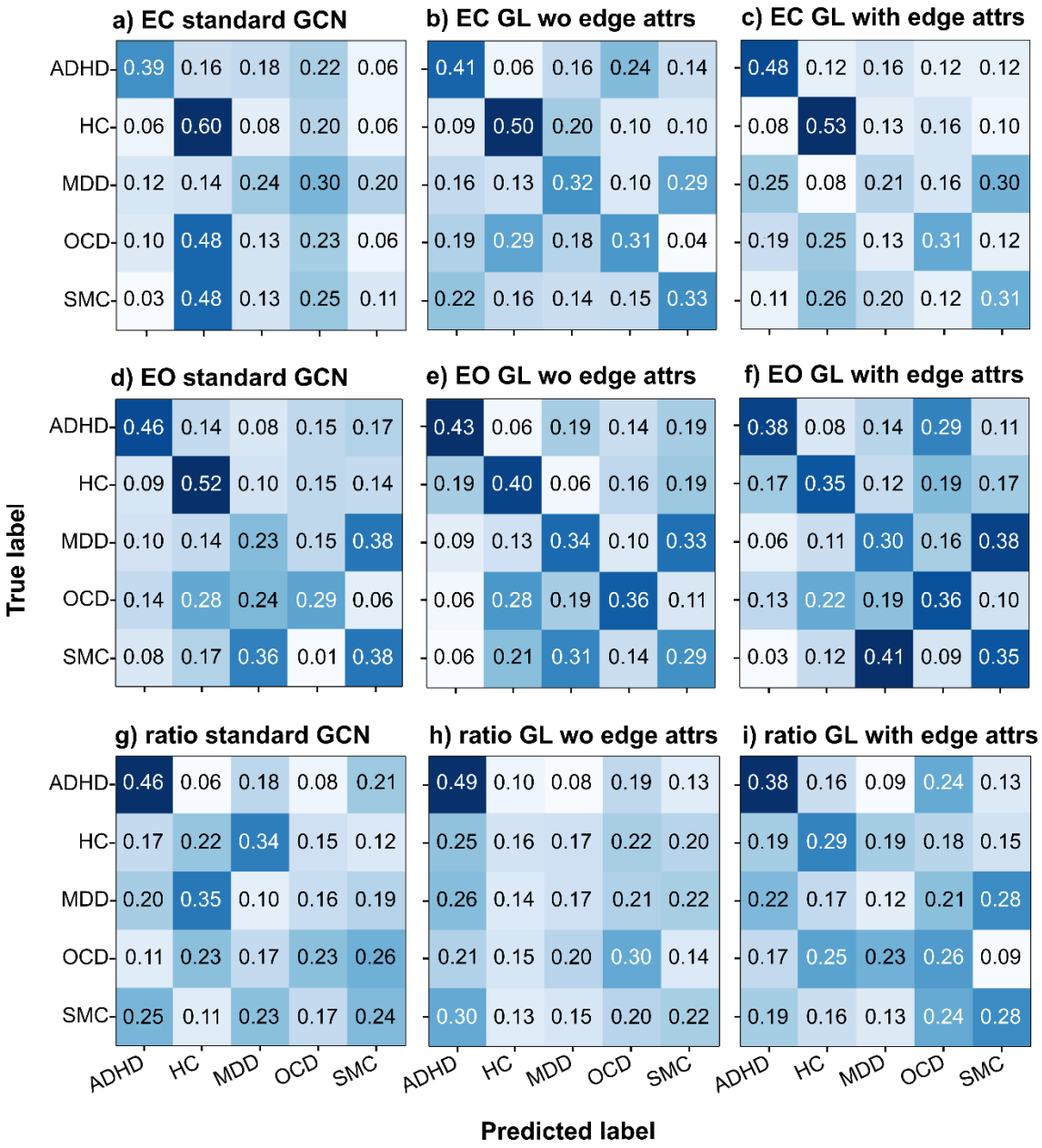


Figure 8 | Confusion matrices from GCN models. The best GCN models for each developed GCN type during tuning, for each feature set and each condition, were used to obtain predictions on the held-out test dataset. The obtained predictions were subsequently visualized in confusion matrices for the standard GCN (a, d, g), the graphLambda without edge attributes (b, e, h), and the graphLambda with edge attributes (c, f, i). 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. For the absolute values see Appendix D. EC = eyes-closed; EO = eyes-open; GCN = Graph Convolutional Network; GL = graphLambda; attrs = attributes; wo = without; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.

The GCNs trained with EC- and EO-derived features, in general, show the highest F1-scores for classifying ADHD, as well as HCs (Table 11). Notably, a pattern of misclassifying MDD with SMC can be observed in the GCNs trained with the EC- and EO-derived features (Figure 8 a-f). The GCNs trained with the ratio-derived features are best able to classify ADHD epochs, however, there seems to be little consistency in the distinction of the other classes (Table 11 & Figure 8 g-i).

4 Discussion

The current study investigated the performance of a multitude of statistical TFR and FC features, derived from different EEG recording conditions. These features were used to train various multiclass ML and GCN models to predict psychiatric status labels in a dataset compromising HCs and four psychiatric disorders: ADHD, MDD, OCD, SMC (per class; 45 participants, 540 epochs). Model performance was compared and inspected to gain insights into the classification of psychiatric disorders and potentially uncover the underlying pathophysiology. It was hypothesized that the inclusion of FC features to the multiclass classification problem would result in an increase in accurate predictions of the four psychiatric disorders and HCs, compared to the models trained with exclusively the statistical TFR features. Additionally, the EC- and EO-derived feature sets were expected to yield better classification compared to the ratio-derived feature sets, except for certain classes like ADHD, which have shown distinct differences between EC and EO EEG data in previous studies. Finally, we expected that the use of a GCN architecture, as well as edge attributes, would better capture the spatial structure of brain activity and consequently improve prediction accuracy across the five classes compared to the standard ML models.

The results show that the ML models trained with statistical TFR features in conjunction with the FC features achieve better classification of the four psychiatric disorders (ADHD, MDD, OCD, SMC) and HCs, as compared to the ML models trained with solely the statistical TFR features. This finding supports the existing literature on the associations between FC and these disorders (Beare et al., 2017; Furlong et al., 2021; Geffen et al., 2022; González-Madruga et al., 2022, 2022; M. Greicius, 2008; M. D. Greicius et al., 2007; Gürsel et al., 2018; Helm et al., 2018; Kaiser et al., 2015; Peng et al., 2022; Wang et al., 2008).

Moreover, except for the SVMs, it can be seen that the feature sets derived from the EC and EO resting-state EEG recordings result in better model performance than the feature sets derived from the ratio between the EC and EO resting-state EEG recording. Nevertheless, except for the extended SVM, the models trained with the ratio-derived features still demonstrate relatively accurate classification of ADHD. This was expected, as the ratio-derived feature sets should be able to capture the difference in the power of the alpha frequency band in EC and EO resting-state EEG data, also called alpha reactivity, which has been shown to be reduced in individuals with ADHD compared to controls (Fonseca et al., 2013).

Contrary to the hypothesis, the GCN models did not outperform the baseline and extended ML models, despite their ability to utilize the intrinsic spatial structure of the brain activity. Moreover, the inclusion of GAT and GIN layers, and edge attributes in the graphLambda models, did not improve classification over standard GCN models. The relatively poor performance of the GCNs could be explained by the complexity of the models leading to overfitting on the train set. We expect that the GCNs were able to learn complex patterns that were specific to the training data, leading to poor generalization to the validation and test data, as seen from the difference in train and validation F1-score curves during training (Appendix C). This is consistent with the increase in complexity of the graphLambda models, including the use of edge attributes, not leading to better classification compared to the standard GCNs. Attempts to mitigate overfitting, such as tuning learning rate and dropout chance, and employing different learning rate optimizers, were not effective in improving generalizability. Future research could explore the generalizability of graphLambda by training the models with a simplified input compared to the current feature sets, for example, by only including one-dimensional node features and edge attributes. An alternative explanation for the performance of the GCNs is the poor correlation between the class labels and the features. As shown by the cluster analyses, there was no class-specific clustering in the two-dimensional representations of the feature sets, thus, suggesting poor correlation between the class labels and the features. We expect that the inclusion of participants with an indication in the study sample, instead of a formal diagnosis, led to the inclusion of a large amount of data that is unrepresentative for the respective disorders. However, it is important to note that poor correlation between the features and class labels would provide a possible

explanation for the relatively poor performance of all models, and not just the performance of the GCNs relative to the standard ML models.

When analyzing the underlying class predictions of the models, it can be seen that the models generally performed best in distinguishing epochs from participants with ADHD and SMC, followed by epochs from HCs. Moreover, a pattern of misclassifying MDD with SMC can be observed in most ML and GCN models trained with the EC- and EO-derived feature sets. A possible explanation for the accurate classification of ADHD is that, due to it being a neurobiological development disorder, the etiology of ADHD is more grounded in neurobiological dysfunction than for example MDD and OCD, in turn possibly leading to more apparent features in brain activity (Tripp & Wickens, 2009). The same argument can be made for the accurate classification of SMC, which has been linked to prominent neurobiological dysfunction (Benito-León et al., 2010; Blackburn et al., 2014; de Groot et al., 2001; Jorm et al., 2004; Reid & Maclullich, 2006). Conversely, HCs are distinguishable by the absence of neurobiological dysfunction relative to the disorders. An additional explanation for the accurate distinction of ADHD and SMC from the other classes could be the difference in the age distributions between the groups in our current analysis (Table 2). In the current study, the participants with ADHD, on average, were the youngest, while those with SMC were the oldest, with the age distributions of the other classes falling between ADHD and SMC. While age was not included in the model, differences in age could be encoded within the EEG data. For instance, 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; Stacey et al., 2021). Interestingly, the SMC and ADHD groups were the two classes with the highest percentage of participants with only an indication by a clinician, as opposed to other classes which consisted of a larger percentage of formally diagnosed participants (Table 2). This suggests that the data from these groups might be less representative of their respective disorders, possibly introducing more noise and heterogeneity in the data. Despite this, these groups still exhibited better classification performance than the groups with a higher percentage of true patients. Nevertheless, it is expected that including only formally diagnosed patients, granted the sample size is sufficient, would likely improve the classification of ADHD and SMC compared to the current classification performance of these specific classes.

A possible explanation for the frequent misclassification of MDD with SMC could be due to the high comorbidity of the conditions. While the included participants in the current analysis officially did not show any indications for comorbidity, previous literature has shown that MDD is highly associated with SMC (Mendes et al., 2021; Mohn & Rund, 2016; Mowla et al., 2008). To directly deal with comorbidity in psychiatric disorders in future research, a modelling approach could target a continuous marker of symptom severity, rather than a diagnostic category. This can be done in combination with a dimensional diagnostic approach, such as the HiTOP method, in order to facilitate objective classification of disorders with clear boundaries.

It is important to note several limitations of the current study. Firstly, while the study does include formally diagnosed patients, a large part of the sample consists of unstandardized indications provided by a psychologist or general practitioner. Moreover, the percentage of formally diagnosed and indicated participants were not evenly distributed among the sampled classes, leading to possible differences in the amount of heterogeneity and noise in the data per class. Future studies with larger sample sizes of formally diagnosed patients could provide more robust insights into the classification of these psychiatric disorders. Secondly, due to computational restraints, the complex GCN models were not developed using nested CV. If a nested CV approach similar to the one used for the standard ML models was utilized for the GCN models, without an increase in computational capability, the hyperparameter optimization of each model would have taken 15 times as long. Considering, this would have increased the entire development time of the GCN models from two to thirty days, we were unable to utilize nested CV for the GCNs in the time-scope of the current project. Thus, for practical reasons, the GCN models were developed with a single train, validation and test set, potentially biasing the performance. Thirdly, while the current study demonstrates the predictive performance of the FC features in general, it does not provide specific feature importances of features within feature sets. This limits insights into the importance of specific brain activity features, such as activity in specific brain regions, for the classification of psychiatric disorders. Future studies are encouraged to compute feature importances after model development to obtain further insight into the underlying pathophysiology of these disorders. Fourthly, it is important to note that

models were evaluated on an epoch-level, as opposed to the participant-level, thus the predictions of the models at this point do not represent a clinical application. The predictions obtained for each epoch per participant can be aggregated to investigate patient-level classification in future studies, for example by classifying the participant with the majority prediction of the epochs corresponding to that individual. Lastly, while the data-driven approach in the current study provides an objective diagnostic methodology, inherent within-disorder heterogeneity and between-disorder homogeneity remain significant challenges due to the traditional psychiatric disorders being defined by symptom complexes with unclear boundaries. We suggest that future studies explore integrating objective data-driven approaches with dimensional alternatives to the traditional nosology, such as the HiTOP method. This could compound the advantages of a data-driven ML approach with dimensional classification, by minimizing within-disorder heterogeneity and between-disorder homogeneity with the HiTOP method, and by the objective diagnosis with the use of biomarkers with a data-driven ML approach.

Nonetheless, a strong point of the current study is the inclusion of a broad spectra of psychiatric disorders, feature sets, EEG conditions, and models, providing a broad perspective on the classification of psychiatric disorders, in turn, providing insights into the pathophysiology. An additional strength is the prevention of data leakage through multiple measures. These include using a stratified grouping K-fold splitting approach for all models, as well as by developing all standard ML models with nested CV. The stratified grouping K-fold splitting approach prevented participant data from appearing in the training set and testing set within the same split, as well as retaining balanced classes, preventing data leakage. Furthermore, considering the nested CV limited data leakage while concurrently training, tuning and evaluating the model on multiple different data splits, the obtained standard ML model performance estimates were less biased as well.

Considering the model performance and the limitations and strengths of the current study, we infer that the presented methodology is not yet suitable for clinical diagnosis of psychiatric disorders. None of the models were able to reliably predict the included classes with high accuracy, and the methodology lacks essential steps for the diagnosis of individuals, for example due to the lack of patient-level predictions. Nevertheless, the results have shown that FC features of brain activity

leads to an improvement in the ML classification of psychiatric disorders on an epoch-level, especially for ADHD, SMC and HCs. This suggests that with the right methodological improvements and biomarkers, a data-driven ML classification approach could become viable for the diagnosis of psychiatric disorders. Therefore, we believe that with the current study as a foundation in combination with the suggested future directions, an alternative to the current nosology can be developed, in turn, objectively and accurately diagnosing psychiatric patients with disorders with clear boundaries, and providing the right treatment for each individual.

In conclusion, the current study provides a broad perspective on classifying psychiatric disorders using EEG-derived features. As hypothesized, the inclusion of FC features resulted in better prediction of psychiatric disorders, suggesting that the FC in brain activity can be a suitable biomarker for the diagnosis of the included disorders. However, in contrast to our hypothesis, the ability of the GCN architecture to represent the intrinsic graphical structure of brain activity did not lead to better distinction between classes. Nonetheless, investigating the underlying class predictions of the models provided valuable insights into the classification of specific psychiatric disorders, such as the accurate distinction of ADHD and SMC from the other classes. The class-specific results imply that biomarkers in brain activity are most distinct for ADHD and SMC compared to the other disorders and HCs. Overall, these findings provide a valuable perspective on the potential utility of EEG biomarkers in the current diagnosis of psychiatric disorders, as well as provide insights in the underlying psychopathology of these conditions.

5 References

Akalin Acar, Z., Acar, C. E., & Makeig, S. (2016). Simultaneous head tissue conductivity and EEG source location estimation. *NeuroImage*, *124*(Pt A), 168–180. https://doi.org/10.1016/j.neuroimage.2015.08.032

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

Alnagger, N., Cardone, P., Martial, C., Laureys, S., Annen, J., & Gosseries, O. (2023). The current and future contribution of neuroimaging to the understanding of disorders of consciousness. *Presse Medicale (Paris, France: 1983)*, *52*(2), 104163. https://doi.org/10.1016/j.lpm.2022.104163

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). American Psychiatric Association.

Amico, F., Canditiis, D. D., Castiglione, F., Pascarella, A., Venerelli, N., Jennifer, V. F., John, H. Y., & Brophy, J. (2023). A resting state EEG study on depressed persons with suicidal ideation. *IBRO Neuroscience Reports*, *14*, 346–352. https://doi.org/10.1016/j.ibneur.2023.03.012

Angulo-Ruiz, B. Y., Ruiz-Martínez, F. J., Rodríguez-Martínez, E. I., Ionescu, A., Saldaña, D., & Gómez, C. M. (2023). Linear and Non-linear Analyses of EEG in a Group of ASD Children During Resting State Condition. *Brain Topography*, *36*(5), 736–749. https://doi.org/10.1007/s10548-023-00976-7

Aviyente, S., Tootell, A., & Bernat, E. M. (2017). Time-frequency phase-synchrony approaches with ERPs. *International Journal of Psychophysiology*, *111*, 88–97. https://doi.org/10.1016/j.ijpsycho.2016.11.006

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

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

Beare, R., Adamson, C., Bellgrove, M. A., Vilgis, V., Vance, A., Seal, M. L., & Silk, T. J. (2017). Altered structural connectivity in ADHD: A network based analysis. *Brain Imaging and Behavior*, *11*(3), 846–858. https://doi.org/10.1007/s11682-016-9559-9

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

Brody, S., Alon, U., & Yahav, E. (2022). *How Attentive are Graph Attention Networks?* (arXiv:2105.14491; Version 3). arXiv. https://doi.org/10.48550/arXiv.2105.14491

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–198. https://doi.org/10.1038/nrn2575

Calhoun, V. D., Pearlson, G. D., & Sui, J. (2021). Data-driven approaches to neuroimaging biomarkers for neurological and psychiatric disorders: Emerging approaches and examples. *Current Opinion in Neurology*, *34*(4), 469–479. https://doi.org/10.1097/WCO.0000000000000967

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, L. (1995). *Time-Frequency Analysis*. Prentice-Hall.

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

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*, 126. https://doi.org/10.1186/1741-7015-11-126

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

Ebrahimzadeh, E., Saharkhiz, S., Rajabion, L., Oskouei, H. B., Seraji, M., Fayaz, F., Saliminia, S., Sadjadi, S. M., & Soltanian-Zadeh, H. (2022). Simultaneous electroencephalography-functional magnetic resonance imaging for assessment of human brain function. *Frontiers in Systems Neuroscience*, *16*, 934266. https://doi.org/10.3389/fnsys.2022.934266

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

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. *Trends in Cognitive Sciences*, *23*(7), 584–601. https://doi.org/10.1016/j.tics.2019.03.009

Fonseca, L. C., Tedrus, G. M. A. S., Bianchini, M. C., & Silva, T. F. (2013). Electroencephalographic alpha reactivity on opening the eyes in children with attention-deficit hyperactivity disorder. *Clinical EEG and Neuroscience*, *44*(1), 53–57. https://doi.org/10.1177/1550059412445659

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

Furlong, S., Cohen, J. R., Hopfinger, J., Snyder, J., Robertson, M. M., & Sheridan, M. A. (2021). Resting-state EEG connectivity in young children with ADHD. *Journal of Clinical Child and Adolescent Psychology : The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *50*(6), 746–762. https://doi.org/10.1080/15374416.2020.1796680

Geffen, T., Smallwood, J., Finke, C., Olbrich, S., Sjoerds, Z., & Schlagenhauf, F. (2022). Functional connectivity alterations between default mode network and occipital cortex in patients with obsessive-compulsive disorder (OCD). *NeuroImage. Clinical*, *33*, 102915. https://doi.org/10.1016/j.nicl.2021.102915

Gong, L., & Cheng, Q. (2019). Exploiting Edge Features for Graph Neural Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9203–9211. https://doi.org/10.1109/CVPR.2019.00943

González-Madruga, K., Staginnus, M., & Fairchild, G. (2022). Alterations in Structural and Functional Connectivity in ADHD: Implications for Theories of ADHD. *Current Topics in Behavioral Neurosciences*, *57*, 445–481. https://doi.org/10.1007/7854_2022_345

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. *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

Greicius, M. (2008). Resting-state functional connectivity in neuropsychiatric disorders. *Current Opinion in Neurology*, *21*(4), 424. https://doi.org/10.1097/WCO.0b013e328306f2c5

Greicius, M. D., Flores, B. H., Menon, V., Glover, G. H., Solvason, H. B., Kenna, H., Reiss, A. L., & Schatzberg, A. F. (2007). Resting-State Functional Connectivity in Major Depression: Abnormally Increased Contributions from Subgenual Cingulate Cortex and Thalamus. *Biological Psychiatry*, *62*(5), 429–437. https://doi.org/10.1016/j.biopsych.2006.09.020

Gürsel, D. A., Avram, M., Sorg, C., Brandl, F., & Koch, K. (2018). Frontoparietal areas link impairments of large-scale intrinsic brain networks with aberrant fronto-striatal interactions in OCD: A meta-analysis of resting-state functional connectivity. *Neuroscience and Biobehavioral Reviews*, *87*, 151–160. https://doi.org/10.1016/j.neubiorev.2018.01.016

Helm, K., Viol, K., Weiger, T. M., Tass, P. A., Grefkes, C., Del Monte, D., & Schiepek, G. (2018). Neuronal connectivity in major depressive disorder: A systematic review.

Neuropsychiatric Disease and Treatment, *14*, 2715–2737. https://doi.org/10.2147/NDT.S170989

Hilbert, D. (with University of California Libraries). (1912). *Grundzüge einer allgemeinen theorie der linearen integralgleichungen*. Leipzig, B. G. Teubner. http://archive.org/details/grundzugeallg00hilbrich

Hofman, A., Lier, I., Ikram, M. A., Wingerden, M. van, & Luik, A. I. (2023). Uncovering psychiatric phenotypes using unsupervised machine learning: A data-driven symptoms approach. *European Psychiatry*, *66*(1), e27. https://doi.org/10.1192/j.eurpsy.2023.13

Homola, D. (2015, May 8). *BorutaPy*. DanielHomola. https://danielhomola.com/feature%20selection/phd/borutapy-an-all-relevant-feature-selection-method/

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

Iorio-Morin, C., Sarica, C., Elias, G. J. B., Harmsen, I., & Hodaie, M. (2022). Chapter 7—Neuroimaging of psychiatric disorders. In M. F. Chernov, J. A. Rzaev, & R. Martínez-Álvarez (Eds.), *Progress in Brain Research* (Vol. 270, pp. 149–169). Elsevier. https://doi.org/10.1016/bs.pbr.2021.12.001

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

Kaiser, R. H., Andrews-Hanna, J. R., Wager, T. D., & Pizzagalli, D. A. (2015). Large-Scale Network Dysfunction in Major Depressive Disorder: A Meta-analysis of Resting-State Functional Connectivity. *JAMA Psychiatry*, *72*(6), 603–611. https://doi.org/10.1001/jamapsychiatry.2015.0071

Karamzadeh, N., Medvedev, A., Azari, A., Gandjbakhche, A., & Najafizadeh, L. (2013). Capturing dynamic patterns of task-based functional connectivity with EEG. *NeuroImage*, *66*, 311–317. https://doi.org/10.1016/j.neuroimage.2012.10.032

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

Kozak, M. J., & Cuthbert, B. N. (2016). The NIMH Research Domain Criteria Initiative: Background, Issues, and Pragmatics. *Psychophysiology*, *53*(3), 286–297. https://doi.org/10.1111/psyp.12518

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

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.

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. 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. https://www.ncbi.nlm.nih.gov/books/NBK390346/

Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

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. *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. *Depression and Anxiety*, *29*(4), 264–281. https://doi.org/10.1002/da.21891

Mendes, T., Cardoso, S., Guerreiro, M., Maroco, J., Silva, D., Alves, L., Schmand, B., Simões do Couto, F., Figueira, M. L., & de Mendonça, A. (2021). Memory awareness in patients with Major Depressive Disorder. *Journal of Psychiatric Research*, *137*, 411–418. https://doi.org/10.1016/j.jpsychires.2021.03.016

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

Mohn, C., & Rund, B. R. (2016). Neurocognitive profile in major depressive disorders: Relationship to symptom level and subjective memory complaints. *BMC Psychiatry*, *16*, 108. https://doi.org/10.1186/s12888-016-0815-8

Morales, S., & Bowers, M. E. (2022). Time-frequency analysis methods and their application in developmental EEG data. *Developmental Cognitive Neuroscience*, *54*, 101067. https://doi.org/10.1016/j.dcn.2022.101067

Mowla, A., Ashkani, H., Ghanizadeh, A., Dehbozorgi, G. R., Sabayan, B., & Chohedri, A. H. (2008). Do memory complaints represent impaired memory performance in patients with major depressive disorder? *Depression and Anxiety*, *25*(10), E92-96. https://doi.org/10.1002/da.20343

Mqawass, G., & Popov, P. (2024). graphLambda: Fusion Graph Neural Networks for Binding Affinity Prediction. *Journal of Chemical Information and Modeling*, *64*(7), 2323–2330. https://doi.org/10.1021/acs.jcim.3c00771

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

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.

Peng, Z., He, T., Ren, P., Jin, L., Yang, Q., Xu, C., Wen, R., Chen, J., Wei, Z., Verguts, T., & Chen, Q. (2022). Imbalance between the caudate and putamen connectivity in obsessive-compulsive disorder. *NeuroImage. Clinical*, *35*, 103083. https://doi.org/10.1016/j.nicl.2022.103083

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

Sejdic, E., Djurovic, I., & Jiang, J. (2009). Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital Signal Processing*, *19*, 153–183. https://doi.org/10.1016/j.dsp.2007.12.004

Stacey, J. E., Crook-Rumsey, M., Sumich, A., Howard, C. J., Crawford, T., Livne, K., Lenzoni, S., & Badham, S. (2021). Age differences in resting state EEG and their relation to eye movements and cognitive performance. *Neuropsychologia*, *157*, 107887. https://doi.org/10.1016/j.neuropsychologia.2021.107887

Stoica, P. G., & Moses, R. (2005). *Spectral analysis of signals*. Pearson, Prentice Hall.

Suzuki, T., Gu, P., Grove, T. B., Hammond, T., Collins, K. M., Pamidighantam, P., Arnold, P. D., Taylor, S. F., Liu, Y., Gehring, W. J., Hanna, G. L., & Tso, I. F. (2023). Abnormally Enhanced Midfrontal Theta Activity During Response Monitoring in Youths With Obsessive-Compulsive Disorder. *Biological Psychiatry*, *93*(11), 1031–1040. https://doi.org/10.1016/j.biopsych.2022.10.020

Taskesen, E., & Reinders, M. J. T. (2016). 2D Representation of Transcriptomes by t-SNE Exposes Relatedness between Human Tissues. *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 den Heuvel, M. P., & Hulshoff Pol, H. E. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, *20*(8), 519–534. https://doi.org/10.1016/j.euroneuro.2010.03.008

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. *Scientific Data*, *9*(1), Article 1. https://doi.org/10.1038/s41597-022-01409-z

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph Attention Networks*. https://paperswithcode.com/paper/graph-attention-networks

Wang, L., Zhu, C., He, Y., Zang, Y., Cao, Q., Zhang, H., Zhong, Q., & Wang, Y. (2008). Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. *Human Brain Mapping*, *30*(2), 638–649. https://doi.org/10.1002/hbm.20530

World Health Organization. (1992). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. World Health Organization.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). *How Powerful are Graph Neural Networks?* https://paperswithcode.com/paper/how-powerful-are-graph-neural-networks

Yin, G., Chang, Y., Zhao, Y., Liu, C., Yin, M., Fu, Y., Shi, D., Wang, L., Jin, L., Huang, J., Li, D., Niu, Y., Wang, B., & Tan, S. (2023). Automatic recognition of schizophrenia from brain-network features using graph convolutional neural network. *Asian Journal of Psychiatry*, *87*, 103687. https://doi.org/10.1016/j.ajp.2023.103687

Zhang, X., Braun, U., Tost, H., & Bassett, D. S. (2020). Data-Driven Approaches to Neuroimaging Analysis to Enhance Psychiatric Diagnosis and Therapy. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *5*(8), 780–790. https://doi.org/10.1016/j.bpsc.2019.12.015

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: Features selected by Boruta feature selection

Appendix A.1 | Number of selected features in the 'statistical Time-Frequency Representation features' feature set per feature type

Feature type		EC	EO	Ratio
Statistical features:	Mean	45	45	23
	Standard deviation	44	38	12
	Median	45	45	31
	Skewness	10	7	0
	Kurtosis	3	0	0
	Total	147	135	66
Frequency bands:	Delta	26	25	7
	Theta	27	25	16
	Alpha	28	24	6
	Beta	27	27	12
	Gamma	39	34	25
Electrode groups:	L-frontal	15	16	5
	M-frontal	16	15	4
	R-frontal	17	16	6
	L-central	16	16	7
	M-central	16	12	4
	R-central	16	15	7
	L-posterior	17	14	12
	M-posterior	16	15	11
	R-posterior	18	16	10

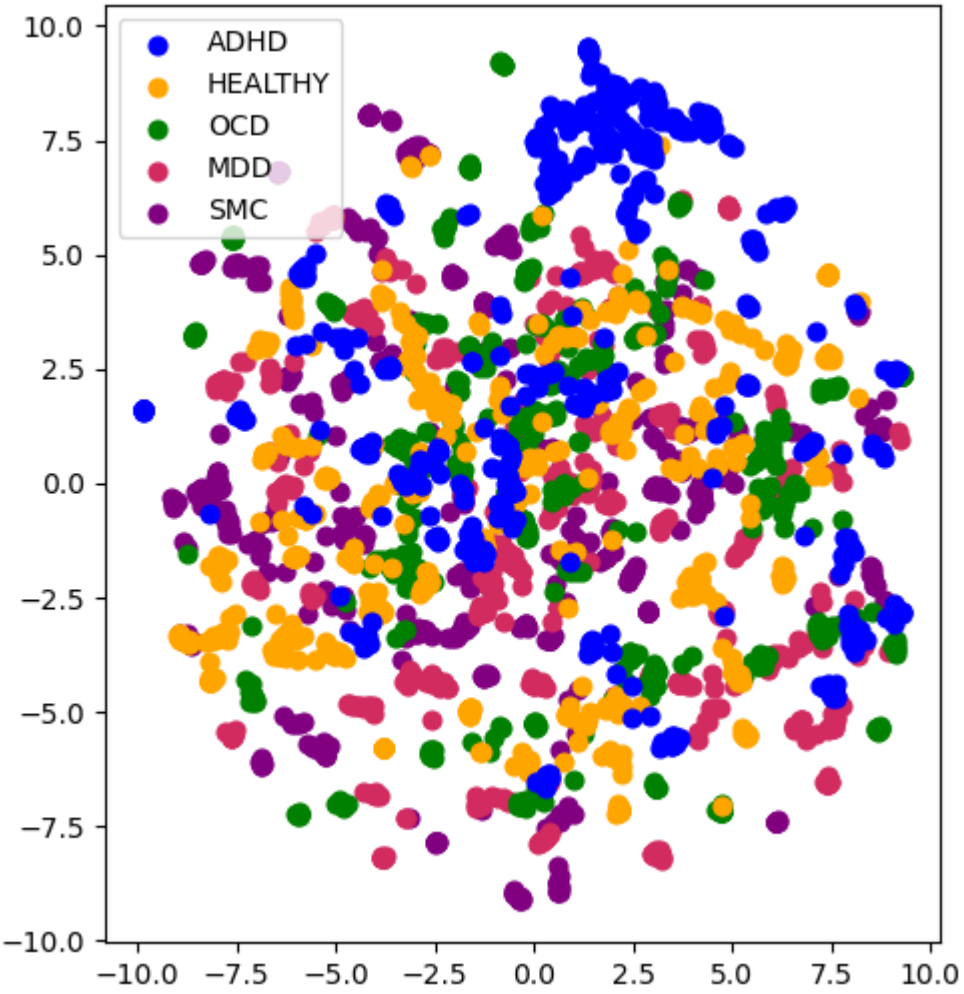
Note. EC = eyes-closed; EO = eyes-open; L = left; M = mid; R = right

Appendix A.2 | Number of selected features in the 'statistical Time-Frequency Representation features + functional connectivity features' feature set per feature type

Feature type		EC	EO	Ratio
FC features:	Total	180	175	103
Statistical features:	Mean	45	45	28
	Standard deviation	45	36	12
	Median	45	45	39
	Skewness	3	0	0
	Kurtosis	1	0	0
	Total	139	126	79
Frequency bands:	Delta	63	54	10
	Theta	63	58	35
	Alpha	63	63	40
	Beta	63	63	34
	Gamma	67	63	63
Electrode groups:	L-frontal	55	52	30
	M-frontal	55	54	30
	R-frontal	55	50	32
	L-central	55	53	31
	M-central	55	53	29
	R-central	55	54	32
	L-posterior	57	55	33
	M-posterior	56	53	34
	R-posterior	56	52	34

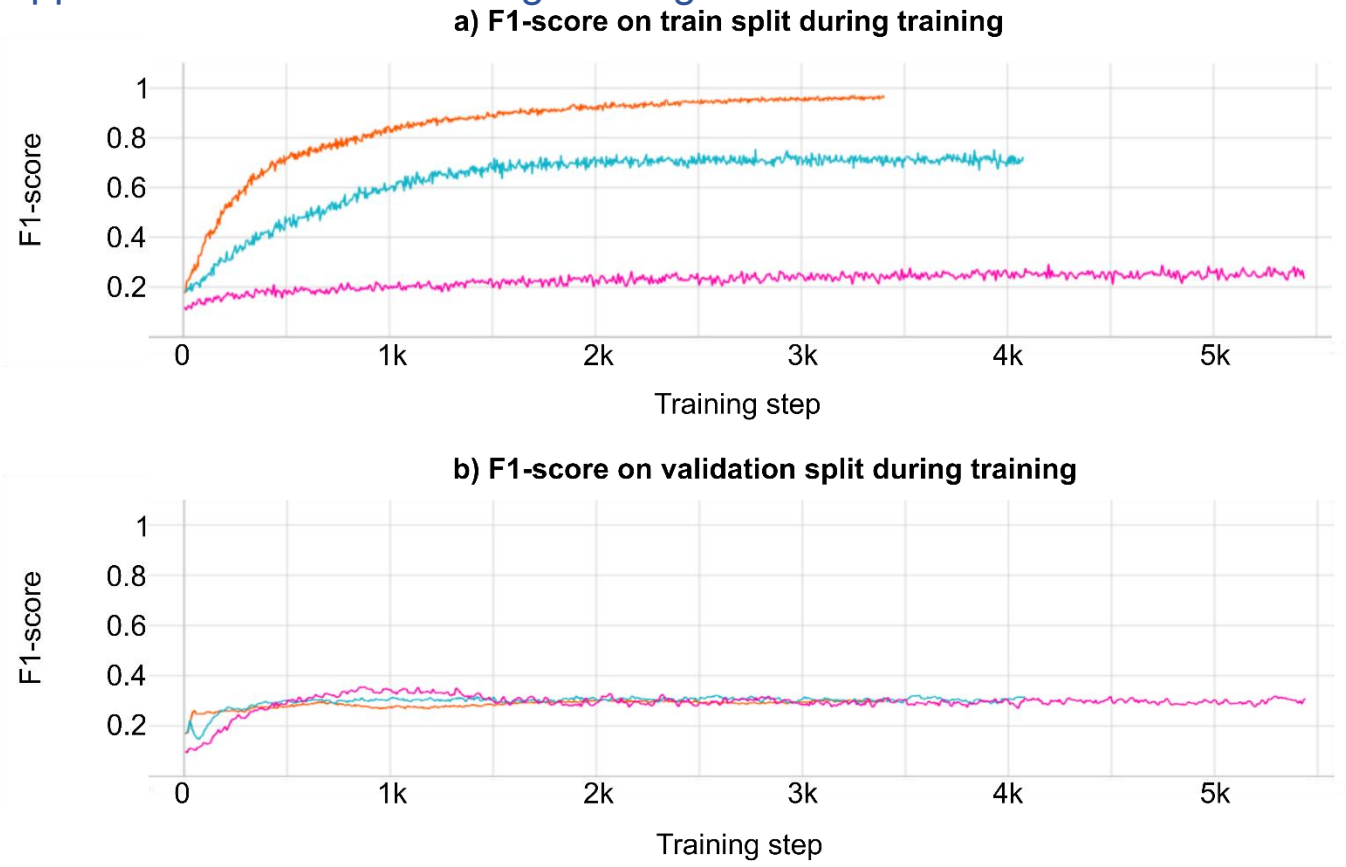
Note. FC = functional connectivity; EC = eyes-closed; EO = eyes-open; L = left; M = mid; R = right

Appendix B: Example visualization of cluster analyses

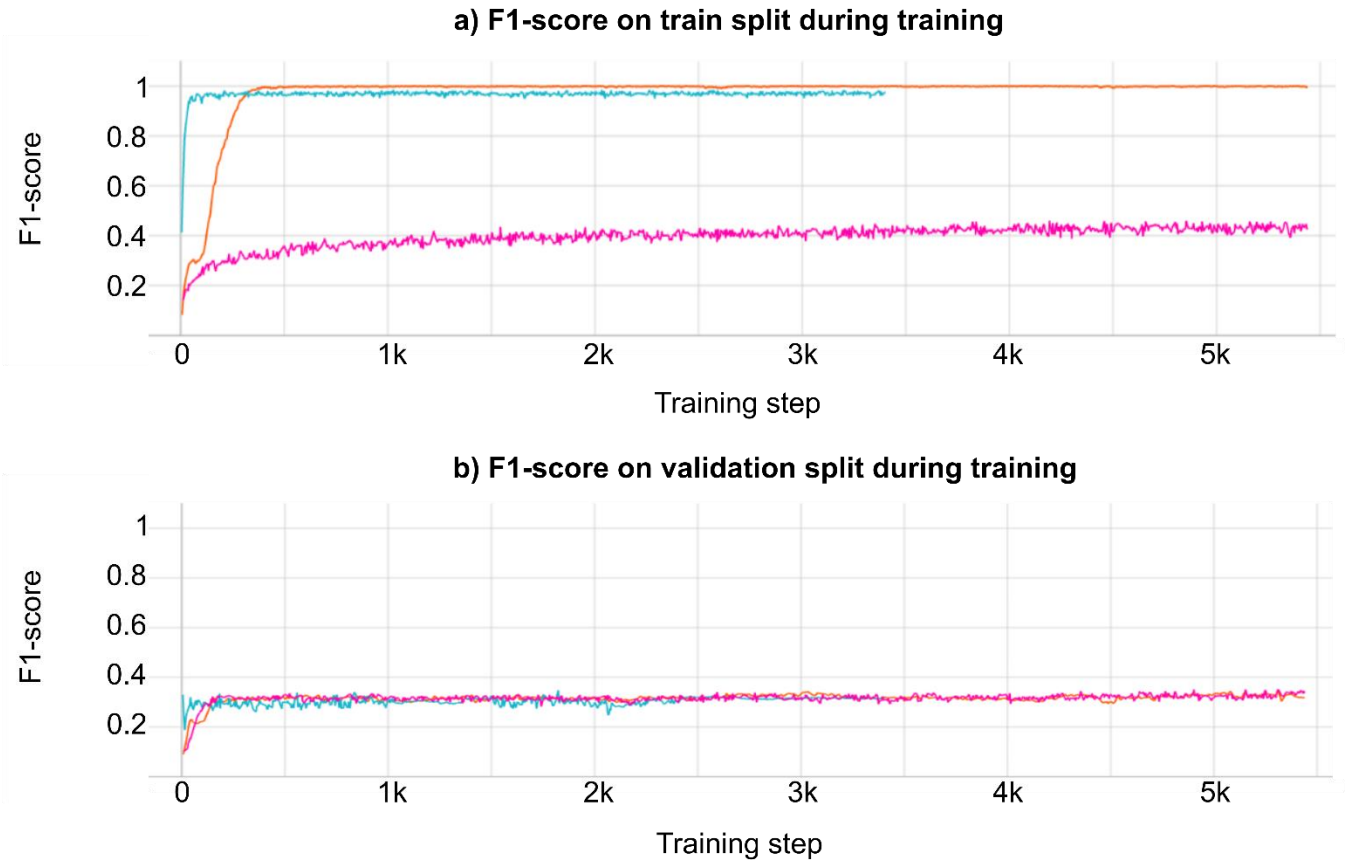


Appendix B.1 | Representative example of t-SNE visualization for all feature sets. Current t-SNE visualization is derived from EC-derived 'statistical Time-Frequency representation features + functional connectivity features' feature set.

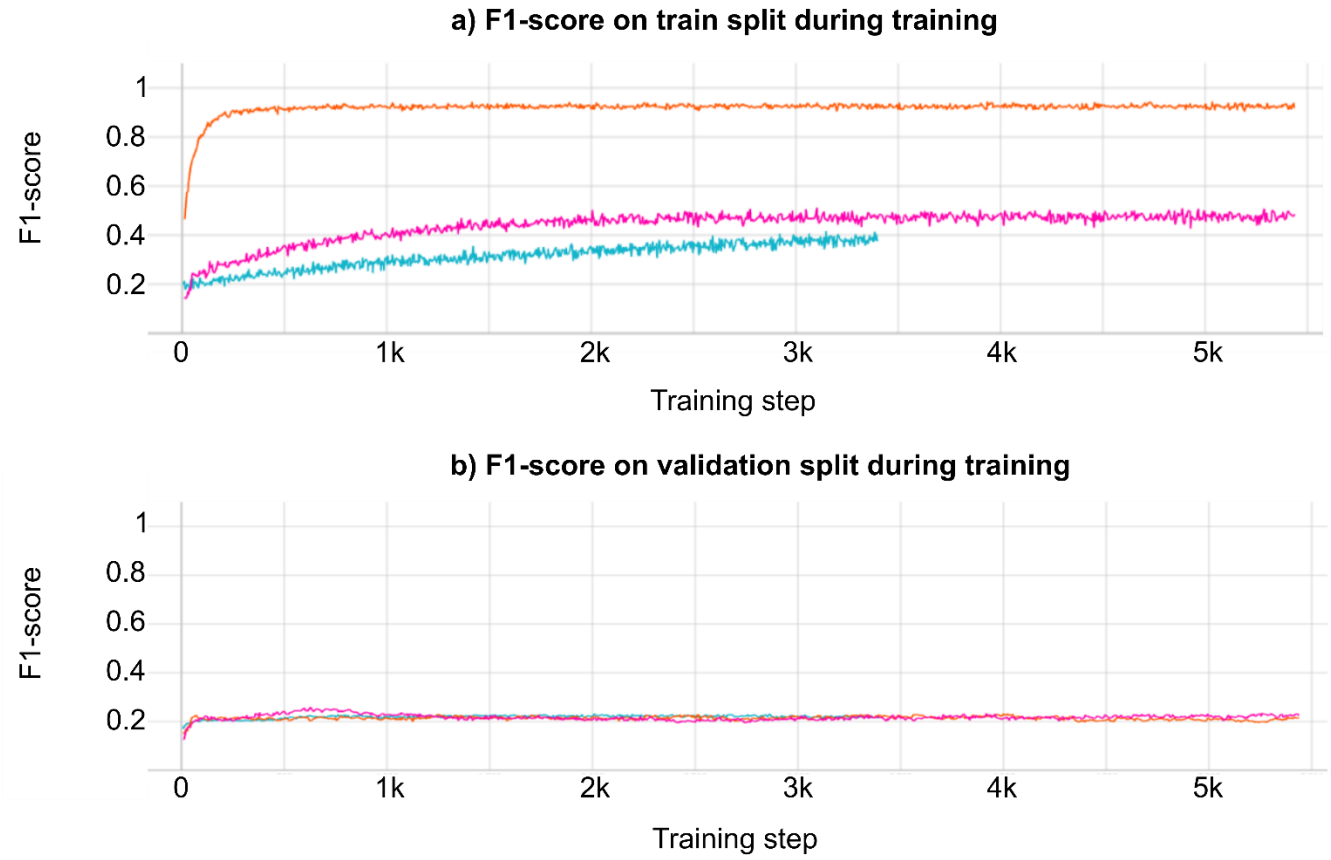
Appendix C: F1-score during training of GCN models



Appendix C.1 | F1-score during training of GCNs with EC-derived features. Showing the F1-score on the train split (a) and the validation split (b), after each training step. Standard GCN (pink), graphLambda without edge attributes (orange), graphLambda with edge attributes (blue).

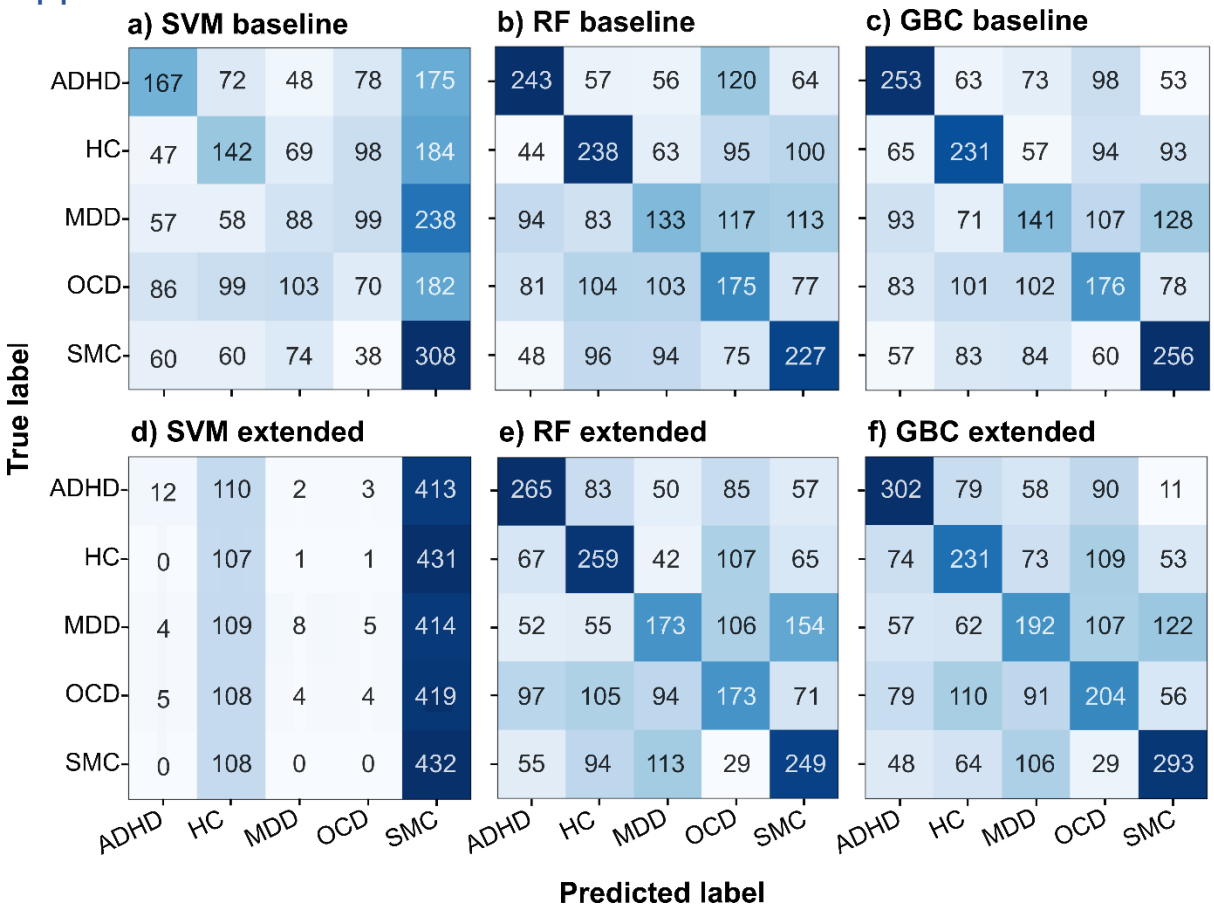


Appendix C.2 | F1-score during training of GCNs with EO-derived features. Showing the F1-score on the train split (a) and the validation split (b), after each training step. Standard GCN (pink), graphLambda without edge attributes (orange), graphLambda with edge attributes (blue).

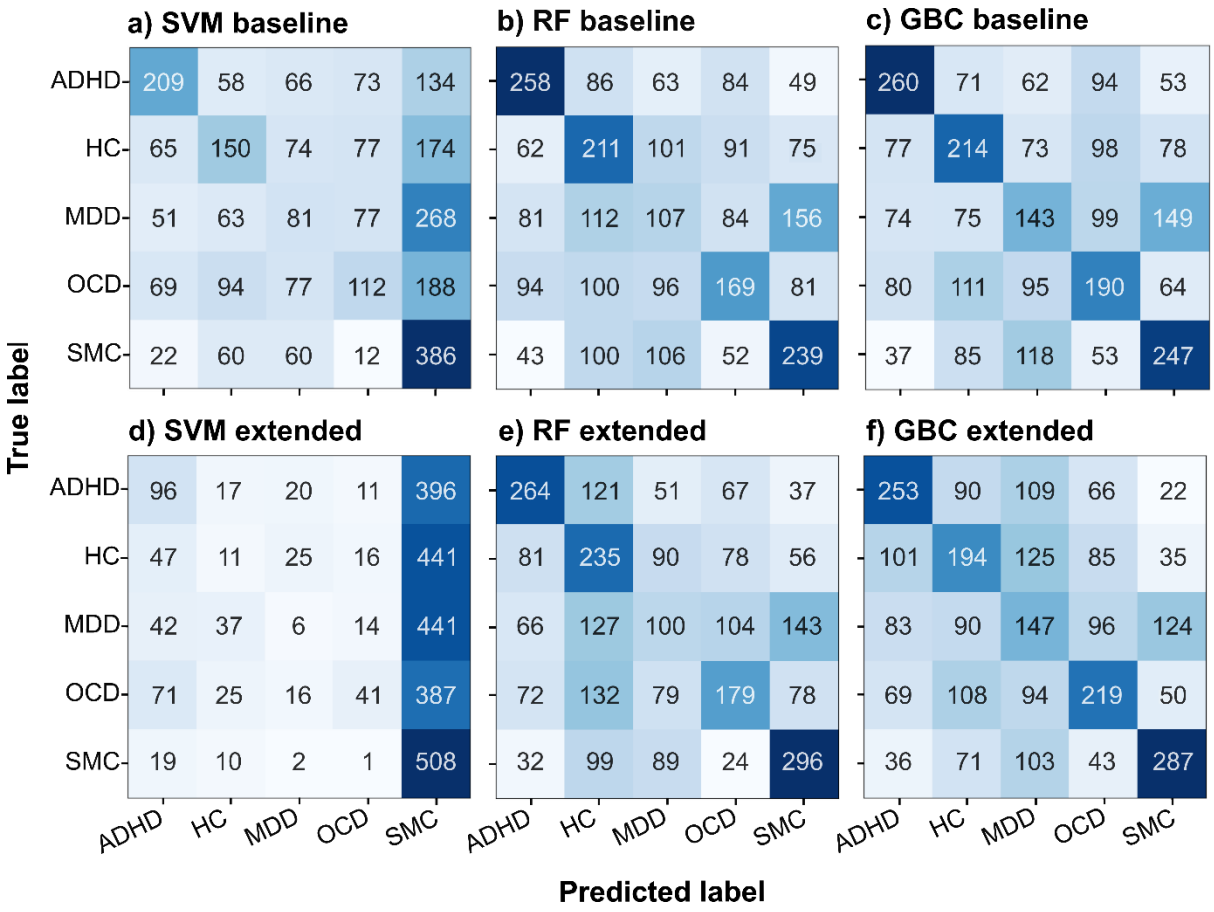


Appendix C.3 | F1-score during training of GCNs with ratio-derived features. Showing the F1-score on the train split (a) and the validation split (b), after each training step. Standard GCN (pink), graphLambda without edge attributes (orange), graphLambda with edge attributes (blue).

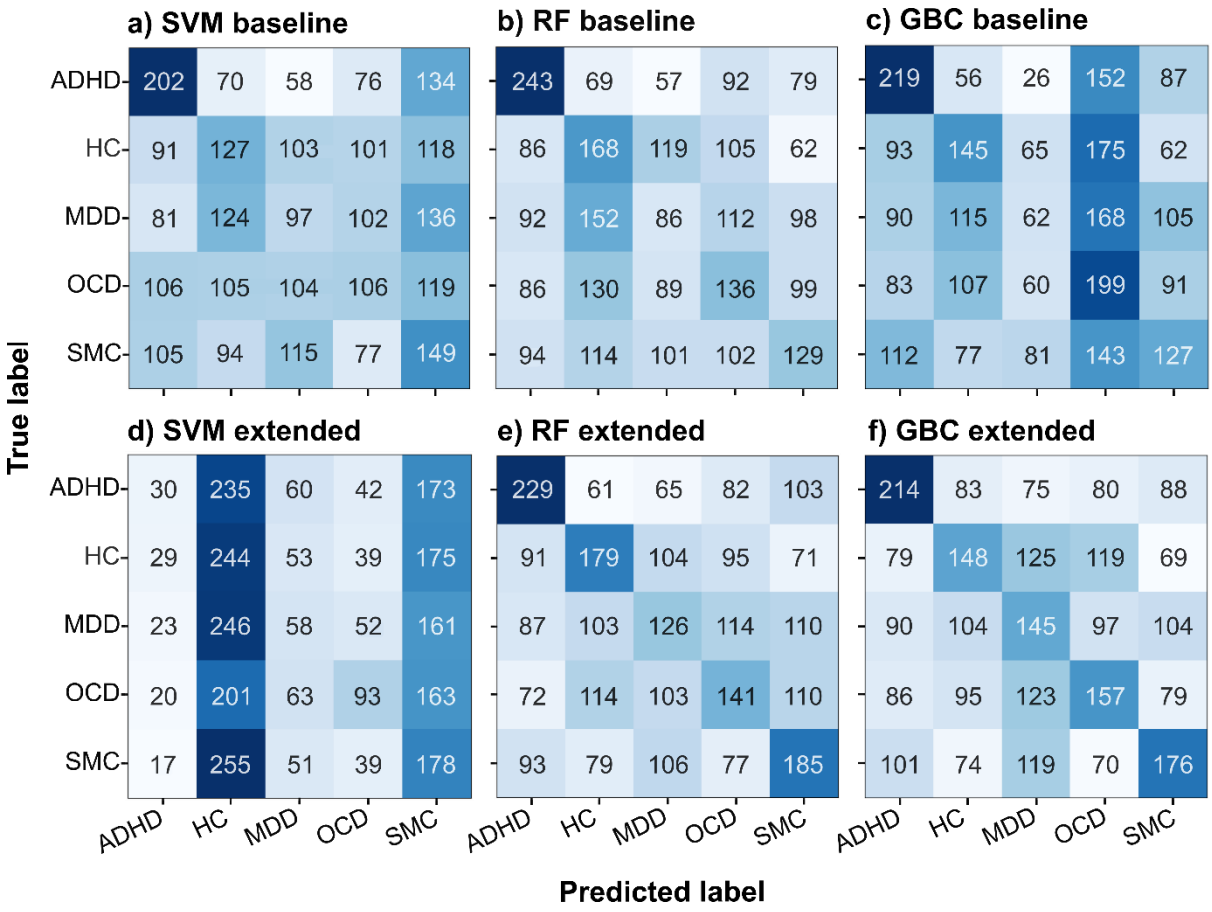
Appendix D: Confusion matrices with absolute values



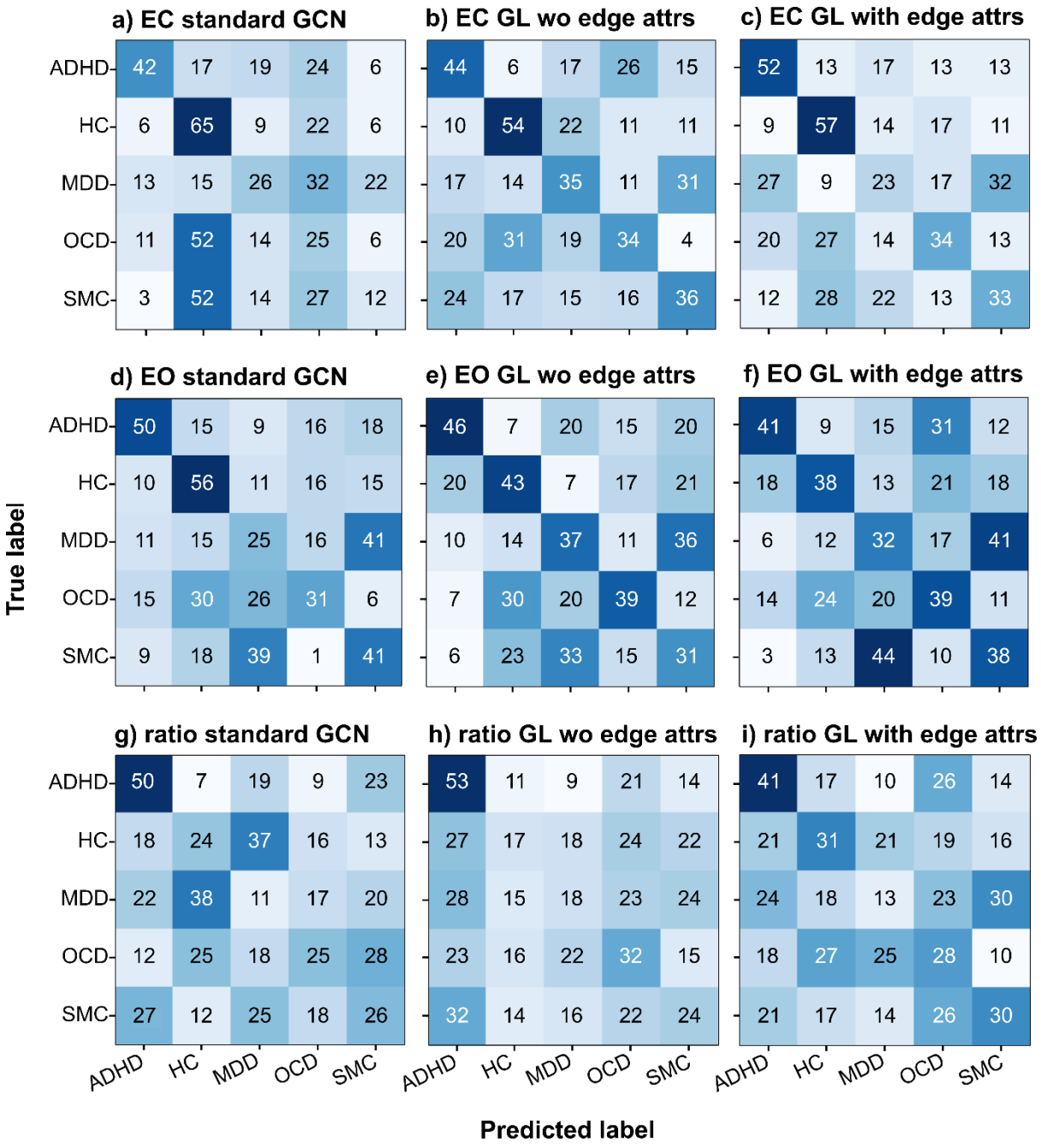
Appendix D.1 | Confusion matrices from ML models trained with EC-derived feature sets. Confusion matrices obtained from a single fold of the nested cross-validation are shown for the baseline models (a – c) and the extended models (d – f), which were trained with the addition of functional connectivity features. Values within each confusion matrix are color-coded to represent the number of times a class was predicted with more predictions visualized as an increasingly darker blue. The range of the color gradient is determined within each individual confusion matrix. SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.



Appendix D.2 | Confusion matrices from ML models trained with EO-derived feature sets. Confusion matrices obtained from a single fold of the nested cross-validation are shown for the baseline models (a – c) and the extended models (d – f), which were trained with the addition of functional connectivity features. Values within each confusion matrix are color-coded to represent the number of times a class was predicted with more predictions visualized as an increasingly darker blue. The range of the color gradient is determined within each individual confusion matrix. SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.



Appendix D.3 | Confusion matrices from ML models trained with ratio-derived feature sets. Confusion matrices obtained from a single fold of the nested cross-validation are shown for the baseline models (a – c) and the extended models (d – f), which were trained with the addition of functional connectivity features. Values within each confusion matrix are color-coded to represent the number of times a class was predicted with more predictions visualized as an increasingly darker blue. The range of the color gradient is determined within each individual confusion matrix. SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.



Appendix D.4 | Confusion matrices from ML models trained with EC-derived feature sets. Confusion matrices obtained from a single fold of the nested cross-validation are shown for the baseline models (a – c) and the extended models (d – f), which were trained with the addition of functional connectivity features. Values within each confusion matrix are color-coded to represent the number of times a class was predicted with more predictions visualized as an increasingly darker blue. The range of the color gradient is determined within each individual confusion matrix. SVM = Support Vector Machine; RF = Random Forest; GBC = Gradient Boosting Classifier; ADHD = Attention Deficit and Hyperactivity Disorder; HC = Healthy Control; MDD = Major Depressive Disorder; OCD = Obsessive Compulsive Disorder; SMC = Subjective Memory Complaints.