
Article:
Schirmer, M.D., Venkataraman, A., Rekik, I. et al. (2021) Neuropsychiatric disease classification using functional connectomics - results of the connectomics in neuroImaging transfer learning challenge. Medical Image Analysis, 70. 101972. ISSN: 1361-8415

https://doi.org/10.1016/j.media.2021.101972

Neuropsychiatric Disease Classification Using Functional Connectomics - Results of the Connectomics in NeuroImaging Transfer Learning Challenge

Markus D. Schirmer[a,b,c,1,*], Archana Venkataraman[d,1], Islem Rekik[e,f,1], Minjeong Kim[g,1], Stewart H. Mostofsky[h,i,j], Mary Beth Nebel[h,i], Keri Rosch[h,j,u], Karen Seymour[h,j], Deana Crocetti[h], Hassna Irzan[l,m], Michael Hütel[m], Sebastien Ourselin[m], Neil Marlow[n], Andrew Melbourne[m,l], Egor Levchenko[o], Shuo Zhou[p], Mwiza Kunda[p], Haiping Lu[p], Nicha C. Dvornek[q,r], Juntang Zhuang[r], Gideon Pinto[s,1], Sandip Samal[s,1], Jennings Zhang[s,1], Jorge L. Bernal-Rusiel[t,1], Rudolph Pienaar[s,u,1], Ai Wern Chung[s,v,1,*]

[a]*Massachusetts General Hospital, Harvard Medical School, Boston, USA*
[b]*German Center for Neurodegenerative Diseases, Bonn, Germany*
[c]*Clinic for Neuroradiology, University Hospital Bonn, Germany*
[d]*Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, USA*
[e]*BASIRA lab, Faculty of Computer and Informatics, Istanbul Technical University, Istanbul, Turkey*
[f]*School of Science and Engineering, Computing, University of Dundee, UK*
[g]*Department of Computer Science, University of North Carolina at Greensboro, USA*
[h]*Center for Neurodevelopmental and Imaging Research, Kennedy Krieger Institute, Baltimore, USA*
[i]*Department of Neurology, Johns Hopkins School of Medicine, USA*
[j]*Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, USA*
[k]*Department of Neuropsychology, Kennedy Krieger Institute, Baltimore, USA*
[l]*Department of Medical Physics and Biomedical Engineering, University College London, UK*
[m]*School of Biomedical Engineering and Imaging Sciences, King's College London, UK*
[n]*Institute for Women's Health, University College London, UK*
[o]*Institute for Cognitive Neuroscience, Higher School of Economics, Moscow, Russia*
[p]*Department of Computer Science, The University of Sheffield, Sheffield, UK*
[q]*Department of Radiology & Biomedical Imaging, Yale University, New Haven, CT, USA*
[r]*Department of Biomedical Engineering, Yale University, New Haven, CT, USA*

[*]Corresponding authors
Email addresses: markus.schirmer@ukbonn.de (Markus D. Schirmer), aiwern.chung@childrens.harvard.edu (Ai Wern Chung)
[1]These authors are the organizers and contributors to the setup of the Challenge.

[s]Fetal-Neonatal Neuroimaging and Developmental Science Center, Division of Newborn Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
[t]Teradyte LLC, Coral Gables, FL, USA
[u]Department of Radiology, Boston Children's Hospital,Harvard Medical School, Boston, MA, USA
[v]Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Abstract

Large, open-source datasets, such as the Human Connectome Project and the Autism Brain Imaging Data Exchange, have spurred the development of new and increasingly powerful machine learning approaches for brain connectomics. However, one key question remains: are we capturing biologically relevant and generalizable information about the brain, or are we simply overfitting to the data? To answer this, we organized a scientific challenge, the Connectomics in NeuroImaging Transfer Learning Challenge (CNI-TLC), held in conjunction with MICCAI 2019. CNI-TLC included two classification tasks: (1) diagnosis of Attention-Deficit/Hyperactivity Disorder (ADHD) within a pre-adolescent cohort; and (2) transference of the ADHD model to a related cohort of Autism Spectrum Disorder (ASD) patients with an ADHD comorbidity. In total, 240 resting-state fMRI (rsfMRI) time series averaged according to three standard parcellation atlases, along with clinical diagnosis, were released for training and validation (120 neurotypical controls and 120 ADHD). We also provided Challenge participants with demographic information of age, sex, IQ, and handedness. The second set of 100 subjects (50 neurotypical controls, 25 ADHD, and 25 ASD with ADHD comorbidity) was used for testing. Classification methodologies were submitted in a standardized format as containerized Docker images through ChRIS, an open-source image analysis platform. Utilizing an inclusive approach, we ranked the methods based on 16 metrics: accuracy, area under the curve, F1-score, false discovery rate, false negative rate, false omission rate, false positive rate, geometric mean, informedness, markedness, Matthew's correlation coefficient, negative predictive value, optimized precision, precision, sensitivity, and specificity. The final rank was calculated using the rank product for each participant across all measures. Furthermore, we assessed the calibration curves of each methodology. Five participants submitted their method

for evaluation, with one outperforming all other methods in both ADHD and ASD classification. However, further improvements are still needed to reach the clinical translation of functional connectomics. We have kept the CNI-TLC open as a publicly available resource for developing and validating new classification methodologies in the field of connectomics.

Keywords: Functional Connectomics, Disease Classification, ADHD, Challenge

1. Introduction

Functional connectomics, or the study of whole-brain synchronization maps, has become of increasing interest to the neuroscientific community in recent years. For example, functional connectomics has provided valuable insight into human cognition (Mill et al. (2017); Shine et al. (2016)), the system-level organization of the brain over development and aging (Somerville et al. (2018); Kaiser (2017)), and whole-brain functional alterations in disease or injury (D'Souza et al. (2019); D'Souza et al. (2020); Venkataraman et al. (2012, 2013b, 2016, 2013a); Ktena et al. (2019); Bonkhoff et al. (2020)). Large, open-source initiatives, such as the Human Connectome Project (Essen et al. (2012)) and the Autism Brain Imaging Data Exchange (Di Martino et al. (2014)), have spurred the development of new and increasingly powerful machine learning strategies to capitalize on these resources.

One popular goal in connectomics is to classify patients from controls. However, objective comparisons of these algorithms across studies can be challenging, due to variations in image acquisition, preprocessing pipeline, and the specific cohort under consideration (Abraham et al. (2017); Pervaiz et al. (2019)). Given these factors, the question of whether a proposed model is capturing biologically relevant and generalizable information about the brain, or simply overfitting to the data, remains to be investigated. In addition to data inconsistencies, performance is assessed using a restricted and non-standardized subset of evaluation metrics, further hindering comparisons across studies (Maier-Hein et al. (2018)). Similar to other fields, scientific challenges provide a way to control for these issues and have been conducted in various domains, such as image registration (Murphy et al. (2011)), lesion segmentation (Heimann et al. (2009); Menze et al. (2014); Mendrik et al. (2015); Commowick et al. (2018); Carass et al. (2017)), and estimation of clinical scores (Wolterink et al. (2016)).

The Connectomics in NeuroImaging Transfer Learning Challenge (CNI-TLC) described here tackles the issues of generalizability and clinical relevance of functional connectomes by leveraging unique resting-state functional MRI (rsfMRI) datasets of Attention-Deficit/Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD), and Neurotypical Controls (NC). ADHD is a chronic neurobehavioral disorder characterized by inattention, hyperactivity, and impulsivity that affects more than 6 million children worldwide (Hamed et al. (2015); Americal Psychiatric Association (2013)). In contrast, ASD patients typically exhibit problems with social skills, communication, and abnormal behavioral habits (Huerta and Lord (2012); Pelphrey et al. (2014); Dowell et al. (2009); McPartland et al. (2011); Americal Psychiatric Association (2013)). While the hallmark behavioral manifestation of ADHD and ASD cohorts differ dramatically, there is a significant comorbidity between the two disorders (Leitner (2014)). Inspired by this finding, we have developed a challenge framework to investigate whether connetome-based features identified in ADHD patients can be transferred to an ASD population for a classification task. Our Challenge setup thereby extends the conventional notion of "transfer learning". As opposed to transferring a previously optimized model for re-training and testing on a new cohort, we propose to transfer and test the learned representation itself, without additional training. This allows us to test whether a robust symptomatology can be learned for ADHD, and if so, whether it can also be extracted in a co-morbid population. Participants were asked to design an ADHD versus NC classification model using rsfMRI time series and demographic measures from 100 and 20 examples of each class for training and validation, respectively. In Task I of the Challenge, we evaluated the classifiers on withheld ADHD and NC data (25 examples of each class). In Task II of the Challenge, we assessed the classification performance of the ADHD model on ASD patients, who have been diagnosed with an ADHD comorbidity (25 examples of each class). Our unique Challenge assesses both the ability of the method to extract functional connectivity patterns related to ADHD symptomatology and how much of this information "transfers" across clinical domains with an overlapping diagnosis. Evaluation was performed using 16 different metrics, while applying cross-validation on the test data, and participants were ranked relative to one another. In this paper, we describe the Challenge data, organization, and detailed evaluation. We also describe the methodology submitted by each participant and the corresponding experimental results.

2. Materials and Methods

2.1. Patient Population

The data used for CNI-TLC were amassed retrospectively across multiple studies conducted by the Center for Neurodevelopmental and Imaging and Research (CNIR) at the Kennedy Krieger Institute (KKI) in Baltimore, MD (see Appendix A for the list of study names and Johns Hopkins IRB approval numbers). The overall cohort includes 145 children diagnosed with ADHD, 25 children with a primary diagnosis of ASD who also meet the diagnostic criteria for ADHD, and 170 NC. All children are between 8-12 years of age and are considered high-functioning based on having a full-scale IQ at or above the normal range; the groups have been matched on age and full-scale IQ. Detailed cohort characteristics are summarized in Table 1.

Table 1: Cohort characterization. For each sample (Training, Validation, and Testing), Controls were matched to patients for all demographics available ($p > 0.05$). FSIQ: Wechsler Intelligence Scale for Children Full Scale Intelligence Quotient; EH: Edinburgh Handedness. SD: Standard Deviation

	All	Training		Validation		Testing			
						Task I) ADHD		Task II) ASD	
		ADHD	Control	ADHD	Control	Patients	Control	Patients	Control
n	340	100	100	20	20	25	25	25	25
Sex (Male; %)	239 (70.3)	70 (70.0)	69 (69.0)	14 (70.0)	14 (70.0)	18 (72.0)	16 (64.0)	17 (68.0)	21 (84.0)
Age (years) (mean (SD))	10.4 (1.3)	10.4 (1.5)	10.3 (1.2)	10.3 (1.4)	10.2 (1.2)	10.3 (1.3)	10.8 (0.9)	10.6 (1.4)	10.7 (1.4)
FSIQ (mean (SD))	111.3 (12.7)	109.2 (12.2)	115.4 (10.4)	104.8 (13.4)	113.7 (14.9)	105.4 (12.3)	111.0 (11.0)	116.2 (14.1)	108.0 (14.9)
EH (mean (SD))	0.7 (0.5)	0.7 (0.5)	0.7 (0.5)	0.7 (0.5)	0.7 (0.4)	0.7 (0.6)	0.8 (0.3)	0.6 (0.7)	0.7 (0.6)

2.2. Clinical Assessment

Participants received an ADHD diagnosis if they met criteria for ADHD using either the Diagnostic Interview for Children and Adolescents (DICA), Fourth Edition (Reich et al. (1997)) or the Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS) for School-Aged Children-Present and Lifetime Version (Kaufman et al. (2000)), in addition to either (1) a t-score of 60 on the Inattentive or Hyperactive subscales of the Conners' Parent or Teacher Rating Scales-Revised Long Version or the Conners-3 (Conners (2002, 2008)), or (2) a score of 2 on at least 6 items on the Inattentive or Hyperactivity/Impulsivity scales of the ADHD Rating Scale-IV, Home or School Versions (DuPaul et al. (1998)). These tests are designed for children

at the age of 6-18 years and were administered by a trained psychologist in CNIR. No additional instructions were given to either the children or examiners for the purposes of this challenge.

Diagnostic criteria for ASD was assessed via two standard instruments: the Autism Diagnostic Observation Schedule Version 2 (ADOS-2) (Gotham et al. (2007)) and the Autism Diagnostic Interview-Revised (ADI-R) (Lord et al. (1994, 2012)). Similar to K-SADS, the ADOS-2 evaluation consists of both structured questions and unstructured narratives. ADOS-2 is used to quantify both socio-communication deficits, as well as repeated/repetitive behaviors. Once again, ADOS-2 was administered by a trained psychologist in CNIR. In contrast, ADI-R is a written questionnaire for parents and/or caregivers about the child. ADI-R provides categorical results for three domains: language and communication, reciprocal social interactions, and repetitive behaviors. The overall ASD diagnosis is based on the summed dimensional scores for each battery.

Inclusion in the NC group required not meeting criteria for any diagnosis on the DICA or K-SADS, having scores below clinical cut-offs on the parent and teacher (when available) Conners' and ADHD Rating Scales, as well as having no immediate family members diagnosed with ADHD or ASD.

In addition to diagnosis, we provided four demographic variables to challenge participants. These variables are age, sex, the Wechsler Intelligence Scale for Children (fourth or fifth editions) Full Scale Intelligence Quotient (FSIQ; Wechsler (2003)), and the Edinburgh Handedness index (Oldfield (1971)).

2.3. Data Acquisition and Preprocessing

The rsfMRI data used in this challenge was acquired on a Philips 3T Achieva scanner housed in the F.M. Kirby Research Center for Functional Brain Imaging at KKI[2]. The acquisition protocol used a single shot, partially parallel gradient-recalled EPI sequence with TR/TE 2500/30ms, flip angle 70°, and voxel resolution $3.05 \times 3.15 \times 3\text{mm}^3$. The scan duration was either 128 or 156 time samples. Children were instructed to relax with their eyes open and focus on a central cross-hair, while remaining still for the duration of the scan. All participants completed a mock scanning session to habituate to the MRI environment.

[2]http://www.kennedykrieger.org/kirby-research-center

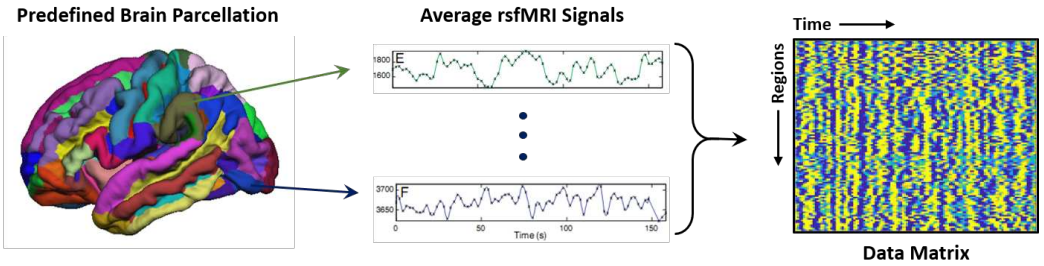

Figure 1: Pipeline of extracting data from a brain parcellation, which was provided for challenge participants.

The rsfMRI data was preprocessed using an in-house pipeline developed by CNIR and implemented in SPM-12 (Friston et al. (2007)). The pipeline included slice timing correction, rigid body realignment, and normalization to the EPI version of the MNI template. The time courses were temporally detrended in order to remove gradual trends in the data. CompCorr (Behzadi et al. (2007)) was performed to estimate and remove spatially coherent noise from the white matter and ventricles, along with the linearly detrended versions of the six rigid body realignment parameters and their first derivatives. From here, the data was spatially smoothed with a 6mm FWHM Gaussian kernel and bandpass filtered between 0.01-0.1 Hz. Finally, spike correction was performed using the AFNI package (Cox (1996)), as an alternative to motion scrubbing.

We released average region-wise time series data for the CNI-TL Challenge, from which participants could compute both static and dynamic connectivity measures. These average signals were computed based on three standard parcellations: (1) the AAL atlas (Tzourio-Mazoyer et al. (2002)), which consists of 90 cortical/subcortical regions and 26 cerebellar regions, (2) the Harvard-Oxford atlas (Desikan et al. (2006)), which consists of 110 cerebral and cerebellar regions, and (3) the Craddock 200 atlas (Craddock et al. (2012)), which is a finer parcellation of 200 regions. The choice of atlases enabled participants to analyze the rsfMRI data at multiple spatial scales. Fig. 1 illustrates our postprocessing workflow for a single parcellation. The mean time courses for each parcellation were aggregated into a single data file for participants to use at their discretion.

2.4. Challenge Design

2.4.1. Data

Submissions to CNI-TLC were evaluated based on two classification tasks: *Task I* - Primary Classification of ADHD versus NC; and *Task II* - Transference Classification of ASD versus NC. The cohort was divided into training (100 ADHD, 100 NC), validation (20 ADHD, 20 NC) and testing datasets (Task I: 25 ADHD, 25 NC; Task II: 25 ASD, 25 NC). The training and validation datasets only consisted of ADHD and NC subjects, in fulfillment of Task I, and was made available for participants to download via the challenge website[3]. Training data was released in June 2019, and validation data followed thirty days later. Each dataset was organized into top directories, one for each subject. The subject directories included four .csv files, one containing the demographic variables, and a separate file for each parcellation. Testing datasets for both Task I and Task II (composed of NC, ADHD and ASD patients with ADHD comorbidity) were not released to the public. The patient and NC cohorts were matched on age, sex, FSIQ, and handedness in each of the three datasets.

2.4.2. Submission Infrastructure

Participants were instructed to submit their trained models as Docker images[4]. Docker is an emerging Platform-As-A-Service (PaaS) product that provides a simple mechanism for bundling applications with all their required dependencies in easily isolated components called containers. In this manner, Docker containerized applications can be executed on all the major computing platforms (Linux, Mac, and Windows) with no additional software requirements besides Docker itself. By using Docker containers as the deployment vector for the trained models, the problem of actually running these models on a single evaluation system was addressed.

For our Challenge, participants were directed to clone a GitHub repository[5]. This repository contained all the basic skeleton components for building a Docker image solution. In addition to the Docker components, a dummy stub Python program, `pl-cni_challenge.py` was provided as a launching point for a solution. Participants could either code directly

[3] http://www.brainconnectivity.net

[4] http://www.docker.com

[5] https://github.com/aichung/pl-cni_challenge

into `pl-cni_challenge.py`, import their existing code as a Python module, or include a suitably compiled executable that can be called using the `os.system()` function in Python.

The repository itself was created by running a cookie-cutter template code[6] that creates plug-ins for ChRIS[7]. The use of the ChRIS system had no significant impact on participants other than defining a standard command line interface contract.

All submissions were required to accept two positional arguments – an `inputDirectory` that would contain the data, and an `outputDirectory` that would store their model predictions. In addition, submitted solutions were to expect the input directory to contain test data with the same folder structure as the released training and validation data, with subject diagnosis information excluded. Participants were instructed to write two text files into the output directory: `classification.txt` containing binary classification labels for each subject, and `score.txt` containing the probability score of each corresponding label in `classification.txt`. Full instructions on how to create, compile and execute a compatible Docker image for submission were provided on the Challenge website. The Docker images were executed on each test subject, one at a time. For consistency, all submissions were evaluated on the same machine.

2.5. Participants

Five solutions were submitted before the deadline of the Challenge. A brief summary of each method is given below. Additional details for each method can be found in Appendix B.

Submission 1 (S1). *MeInternational.* The method is based on building classification pipelines by using permutations of functional connectivity matrices, anatomical atlases, and classification algorithms (Abraham et al. (2017)). From each atlas and time series, correlation, covariance, partial correlation, precision, and tangent embedding for functional connectivity were estimated. Classification was based on support vector machines (SVMs), linear regression (l_1 or l_2 regularization), random forest, k-nearest neighbor, and naive Bayes classifiers. Each combination was tested on the training data and evaluated on the validation data set, where the best performing pipeline was chosen based on its prediction accuracy score.

[6] https://github.com/FNNDSC/cookiecutter-chrisapp
[7] https://chrisproject.org

Submission 2 (S2). *HSE.* The core principle of this submission was the utilization of eigenvalues of the normalized Laplacian. In brief, for each connectome, the normalized Laplacian and its corresponding eigenvalues were calculated. The full set of eigenvalues were subsequently used as features in an SVM classification algorithm using a polynomial kernel.

Submission 3 (S3). *ShefML.* This solution consists of two stages: First, pairwise region of interest (ROI) features, ROI-to-ROI, were extracted from the rsfMRI time series by computing Tangent Pearson connectivity (Kunda et al. (2020)). Then, an SVM was trained which is regularized by the statistical independence (Gretton et al. (2005)) between the classifier decision scores and three types of demographic information: gender, age, and handedness score (Zhou et al. (2020)).

Submission 4 (S4). *ShefML.* In this approach, five types of features were extracted from the time series by computing: mean and standard deviation, Pearson correlation, Tangent (Varoquaux et al. (2010a)), covariance (Varoquaux et al. (2010b)), and Tangent Pearson (Kunda et al. (2020)) connectivity. Subsequently, five classifiers (Zhou et al. (2020)), one per feature, were trained and classifications were combined by majority voting.

Submission 5 (S5). *YaleIPAG.* Here, an LSTM-based (long short-term memory) network was used to learn directly from time-series data (Dvornek et al. (2017)). Twenty-two AAL ROIs were first selected based on consistent connectivity differences between ADHD and controls in bootstrapped samples. The time-series from these ROIs were input to an LSTM, with the demographic data used in hidden and cell state initialization (Dvornek et al. (2018)).

2.6. Evaluation and Ranking

2.6.1. Challenge Evaluation Metrics

Taking an inclusive approach, we assessed multiple measures commonly used in classification tasks, such as accuracy and area under the curve (AUC), along with distributional measures, such as geometric-mean and optimized precision. This approach provides an intuitive and robust characterization of each submission. A full list of the utilized measures and their interpretation is given in Table 2, while a detailed description of these measures can be found elsewhere in the literature (Hossin and Sulaiman (2015)).

Table 2: Summary of evaluation metrics, including abbreviation (Abbr.) used, a short description, and condition under which one participant outperforms another ('Better if').

Measure	Abbr.	Description	Better if
Accuracy	Acc	Ratio of correct predictions over the total number of instances evaluated	higher
Area under curve	AUC	Reflects overall ranking performance of classifier	higher
F1-score	F1	Harmonic mean between recall (sensitivity) and precision	higher
False discovery rate	FDR	Fraction of misclassified positive samples in relation to the number of the total positive classified samples	lower
False negative rate	FNR	Also known as miss rate. Fraction of misclassified negative samples in relation to the number of positive samples	lower
False omission rate	FOR	Fraction of misclassified negative samples in relation to the number of the total negative classified samples	lower
False positive rate	FPR	Fraction of misclassified positive samples in relation to the number of total negative classified samples	lower
Geometric mean	GM	Geometric mean of sensitivity and precision	higher
Informedness	Inf	Also known as Youden's J statistic. It summarizes the true positive and true negative rates of a classifier	higher
Markedness	Mark	Summarizes the positive (prediction) and negative predictive value of a classifier	higher
Matthews correlation coefficient	MCC	Correlation coefficient between the observed and predicted binary classification with values between -1 (total disagreement) and 1 (perfect prediction)	higher
Negative predictive value	NPV	Fraction of correctly classified negative samples in relation to the number of the total negative classified samples	higher
Optimized precision	OP	Measure aiming to simultaneously minimize the difference in sensitivity and specificity, while maximizing their sum. Precision is subsequently "corrected" by the ratio of this difference and sum.	higher
Precision	Pre	Fraction of correctly classified positive samples in relation to the number of the total positive classified samples	higher
Sensitivity	Sen	Fraction of positive samples that are correctly classified	higher
Specificity	Spec	Fraction of negative samples that are correctly classified	higher

2.6.2. Comparison and Ranking Strategy

We first evaluated the model performances on the validation set to assess the primary classification task of ADHD versus NC. This evaluation provides a baseline measure of performance for each algorithm on data that has been made available to the participants. It also allows us to investigate the drop in performance when presenting the classifier with unseen test data later on.

We utilized a 5-fold cross-validation scheme for statistical testing of each submission on the unseen data. Namely, we "randomly" divided our testing dataset into five equal sized, statistically indistinguishable (in terms of sex, age, FSIQ, and handedness), disjointed folds[8] and calculated the evaluation metrics (see Table 2), using data from four of the five folds. This process yields five quantitative values for each evaluation metric. The cross validation procedure was repeated 100 times (on different random splits of the data),

[8]https://github.com/mdschirmer/MDS;Schirmer et al. (2019)

resulting in a distribution of 500 values for each metric.

Our initial ranking was based on the median of the distributions, where the ranking between participants was statistically evaluated using a pairwise Wilcoxon test. Considering the total of 320 tests (ten comparisons between challengers with 16 measures each for two classification tasks, i.e., ADHD and ASD), we set the significance level to $p \leq 0.0001$ (Bonferroni correction of 0.05/320). Finally, we calculated the rank product (geometric mean) for each participant, which gave us the final ranking of submissions.

In addition, we evaluated submissions based on calibration curves. With calibration curves, we can gain an idea of the model's behavior and confidence in performing the classification tasks. These curves assess if the model can reliably estimate the probability of the diagnosis by plotting the mean predicted probability against the true probability for each user-specified probability bin. Here, we summarize the results of the cross-validation approach as a single calibration curve using 10 bins of width 0.1, and fit a linear model to the predicted probability against the observed probability for comparison.

3. Results

The patient cohort utilized in this challenge was on average 10.42 years old, 70.3% male, with an average FSIQ of 111.3, and a handedness score of 0.7. Neurotypical controls in training, validation, and test data were not significantly different from their corresponding patient population in terms of sex, age, FSIQ, and handedness (all $p > 0.05$).

The evaluation of the performance for all participants and datasets is summarized in Table 3. Using the validation set performance as a baseline for the Primary Classification Task I (ADHD), we observed a large drop in performance for participants S1, S3, and S4, across all metrics against the test set. The performance of the methods submitted by participants S2 and S5, however, remained relatively stable. For the Transference Classification Task II (ASD with ADHD comorbidity), we see a further decrease in performance compared with Task I, across all metrics and for all participants.

Table 3: Summary of performance for each participant (S1 to S5) given as the median of each evaluation metric for validation, Task I, and Task II test datasets.

	Task I - ADHD										**Task II - ASD**				
	Validation					**Test**					**Test**				
	S1	**S2**	**S3**	**S4**	**S5**	**S1**	**S2**	**S3**	**S4**	**S5**	**S1**	**S2**	**S3**	**S4**	**S5**
Acc	0.75	0.53	0.73	0.83	0.68	0.5	0.55	0.53	0.55	0.68	0.45	0.53	0.45	0.43	0.53
AUC	0.73	0.62	0.85	0.89	0.72	0.48	0.63	0.54	0.47	0.66	0.45	0.48	0.43	0.41	0.56
F1	0.79	0.56	0.73	0.82	0.68	0.6	0.57	0.51	0.54	0.7	0.55	0.54	0.43	0.41	0.5
FDR	0.32	0.48	0.29	0.16	0.33	0.5	0.45	0.48	0.44	0.33	0.54	0.48	0.55	0.58	0.47
FNR	0.05	0.4	0.25	0.2	0.3	0.25	0.4	0.5	0.5	0.3	0.3	0.45	0.6	0.6	0.5
FOR	0.08	0.47	0.26	0.19	0.32	0.5	0.44	0.48	0.45	0.3	0.62	0.47	0.54	0.58	0.48
FPR	0.45	0.55	0.3	0.15	0.35	0.75	0.5	0.5	0.4	0.35	0.8	0.5	0.5	0.55	0.45
GM	0.8	0.56	0.73	0.82	0.68	0.61	0.57	0.51	0.54	0.7	0.56	0.54	0.43	0.41	0.5
Inf	0.5	0.05	0.45	0.65	0.35	0.00	0.1	0.05	0.1	0.35	-0.1	0.05	-0.1	-0.15	0.05
Mark	0.6	0.05	0.45	0.65	0.35	0.00	0.1	0.05	0.1	0.36	-0.16	0.05	-0.1	-0.15	0.05
MCC	0.55	0.05	0.45	0.65	0.35	0.00	0.1	0.05	0.1	0.35	-0.13	0.05	-0.1	-0.15	0.05
NPV	0.92	0.53	0.74	0.81	0.68	0.5	0.56	0.52	0.55	0.7	0.39	0.53	0.46	0.42	0.52
OP	29.73	20.86	28.97	32.97	26.96	19.5	22	20.95	22	26.96	17.33	20.86	18	16.94	20.86
Pre	0.68	0.52	0.71	0.84	0.67	0.5	0.55	0.52	0.56	0.67	0.46	0.52	0.45	0.42	0.53
Sen	0.95	0.6	0.75	0.8	0.7	0.75	0.6	0.5	0.5	0.7	0.7	0.55	0.4	0.4	0.5
Spec	0.55	0.45	0.7	0.85	0.65	0.25	0.5	0.5	0.6	0.65	0.2	0.5	0.5	0.45	0.55

Figures 2 and 3 illustrate the distribution of each evaluation metric for each participant's model on Task I and Task II, respectively. Each subplot indicates which of the submissions performs best with respect to a specific metric, as described in Table 2. The performance evaluation was based on the cross-validation setup and significance between ranks are delineated.

Table 4 summarizes the participant rankings based on the results of the metrics presented in Figure 2 and 3, including the best median for each metric (minimum or maximum) across participants. The overall ranking for Task I between ADHD and NC is as follows (participant (rank product)): S5 (1.1), S2 (2.4), S4 (2.7), S1 (3.1), and S3 (3.6). The overall ranking for the Transference Classification Task II between ASD and NC is as follows: S5 (1.3), S2 (1.4), S1 (2.3), S3 (2.6), and S4 (3.6). In both Tasks, S5 and S2 consistently ranked first and second, respectively, in our evaluation.

Table 4: Rankings of each participant for each measure and both classification tasks, including the best median for each metric across classifiers for ADHD and ASD test cohorts.

	Rankings										**Best median metric**	
	Task I - ADHD					**Task II - ASD**					**ADHD**	**ASD**
	#1	**#2**	**#3**	**#4**	**#5**	**#1**	**#2**	**#3**	**#4**	**#5**		
Acc	S5	S2/S4		S3	S1	S2/S5		S3	S1	S4	0.68	0.53
AUC	S5	S2	S3	S1	S4	S5	S2	S1	S3	S4	0.66	0.56
F1	S5	S1	S2	S4	S3	S1	S2	S5	S3	S4	0.70	0.55
FDR	S5	S4	S2	S3	S1	S2/S5		S1/S3		S4	0.33	0.47
FNR	S1	S5	S2	S3/S4		S1	S2	S5	S3/S4		0.25	0.30
FOR	S5	S2	S4	S3	S1	S2/S5		S3	S4	S1	0.30	0.47
FPR	S5	S4	S2/S3		S1	S5	S2	S3	S4	S1	0.35	0.45
GM	S5	S1	S2	S4	S3	S1	S2	S5	S3	S4	0.70	0.56
Inf	S5	S2/S4		S3	S1	S2/S5		S3	S1	S4	0.35	0.05
Mark	S5	S2/S4		S3	S1	S2/S5		S3	S1/S4		0.36	0.05
MCC	S5	S2/S4		S3	S1	S2/S5		S3	S1/S4		0.35	0.05
NPV	S5	S2	S4	S3	S1	S2/S5		S3	S4	S1	0.70	0.53
OP	S5	S2/S4		S3	S1	S2/S5		S3	S1/S4		26.96	20.86
Pre	S5	S4	S2	S3	S1	S2/S5		S1/S3		S4	0.67	0.53
Sen	S1	S5	S2	S3/S4		S1	S2	S5	S3/S4		0.75	0.70
Spec	S5	S4	S2/S3		S1	S5	S3	S2	S4	S1	0.65	0.55

Figure 4 shows the calibration curve for each submission and classification task. A good classification model is represented by a sigmoid or step function. Generally, we observe an expected positive trend, where subjects with higher probability scores are more likely to be ADHD patients in Task I (Figure 4A). For classifying ASD patients with ADHD comorbidity, most methodologies predominantly assigned higher probabilities to controls (Figure 4B). Additionally, out of all submissions, three did not use the entire predicted probability spectrum (S2, S3, S4), with a reversal in their linear fit between Task I and Task II.

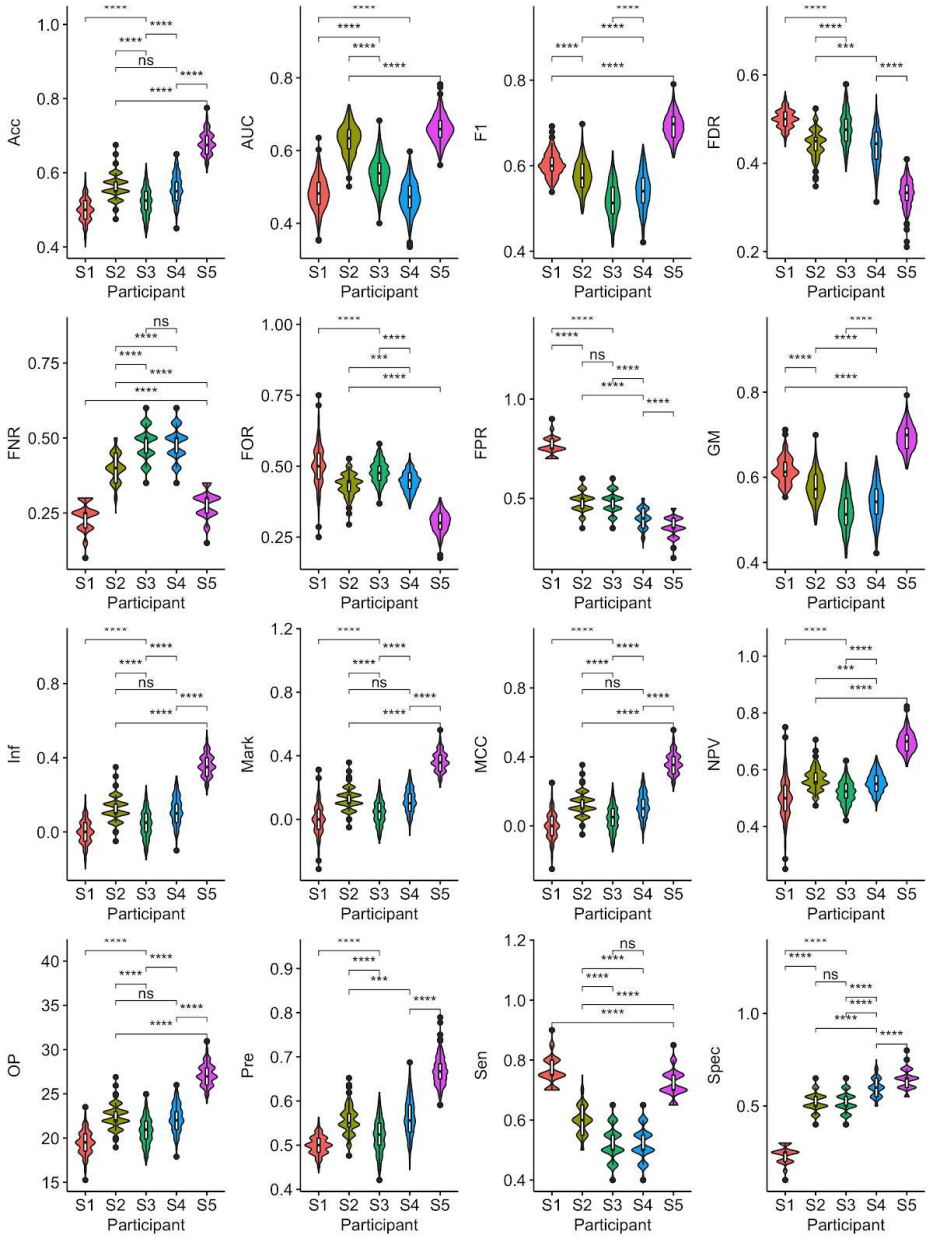

Figure 2: Primary Classification Task I (ADHD versus NC) evaluation measure distributions for each participant. Statistical significance was determined based on pair Wilcoxon test (ns: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; ****: $p < 0.0001$), with a Bonferroni-corrected level of significance at $p \leq 0.0001$.

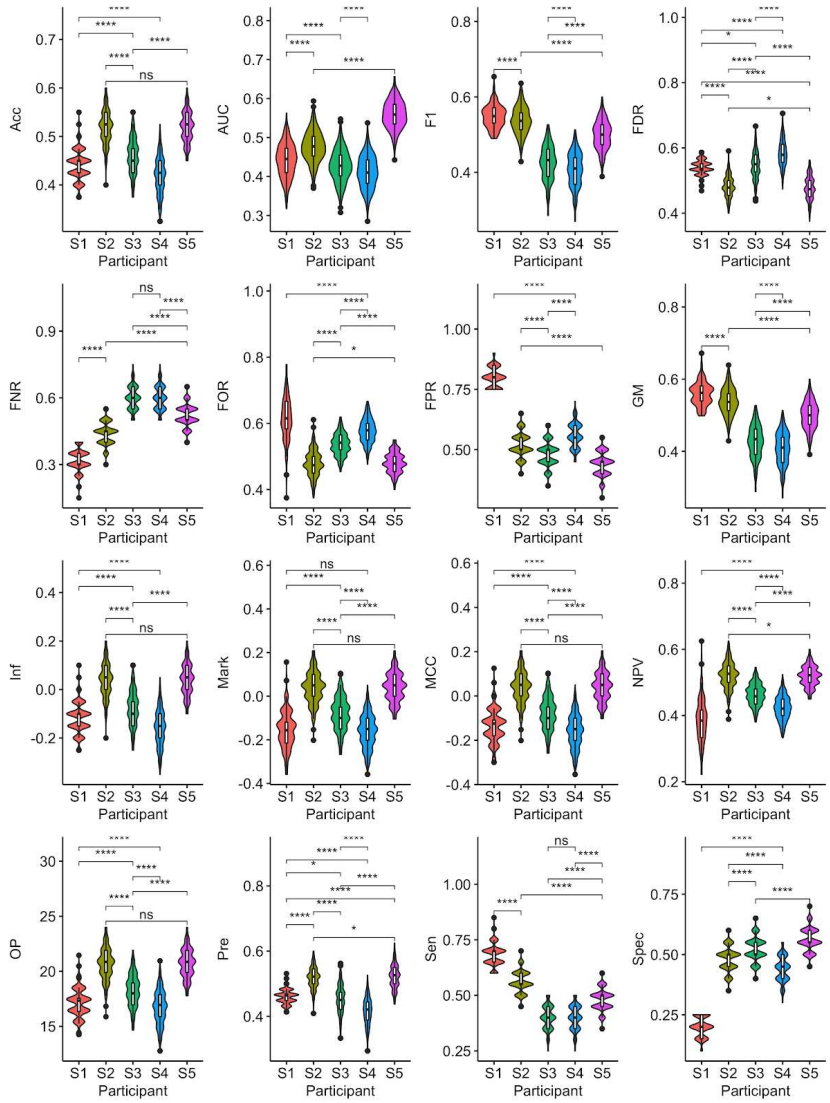

Figure 3: Transference Classification Task II (ASD versus NC) evaluation measure distributions for each participant. Statistical significance was determined based on pair Wilcoxon test (ns: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; ****: $p < 0.0001$), with a Bonferroni-corrected level of significance at $p \leq 0.0001$.

4. Discussion

In this paper, we have described the setup, standardized assessment, and results of the first Connectomics in Neuroimaging Transfer Learning Chal-

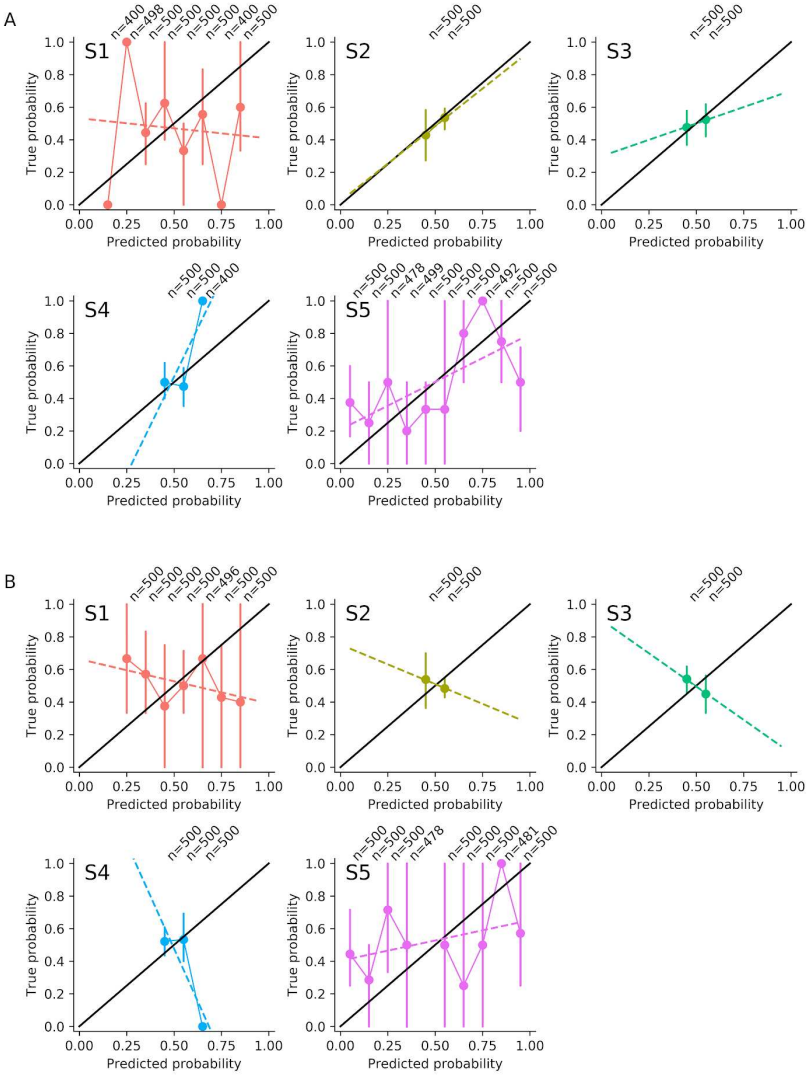

Figure 4: Calibration curves for classification A) Task I (ADHD) and B) Task II (ASD). The n indicates the number of subsets that contributed to the distribution within each bin. Dashed lines represent the resulting linear regression from the data.

lenge, hosted at the 22nd International Conference on Medical Image Computation and Computer Assisted Intervention (MICCAI) 2019 in Shenzhen, China. The CNI-TL Challenge was designed to probe the generalizability and clinical relevance of classification methodologies, which are now growing

in popularity for functional connectivity data.

CNI-TLC combines two developmental disorders, ADHD and ASD, which individually have complex yet distinct behavioral phenotypes and diagnostic assessments. At the same time, the co-occurrence of these disorders is high with many ASD children also exhibiting the stereotypical attention problems and impulsivity of ADHD. In fact, connectomics studies have combined ASD and ADHD populations to disentangle their joint clinical presentations, with findings of both shared and distinct functional network features between ADHD and ASD (Ray et al. (2014); Martino et al. (2013); Lake et al. (2019)). Our unique setup extends the conventional notion of "transfer learning" to further probe this phenomenon. Rather than transferring just the model architecture optimized for one cohort and re-training it on a second cohort, we transfer the learned representations themselves. In this manner, we test whether a robust symptomatology can be learned for ADHD, and if so, whether it can also be extracted in a co-morbid population. Furthermore, our Challenge setup is in line with recent trends in computational neuroscience to develop methods that learn from multiple populations, so as to increase their clinical relevance (Douw et al. (2019)).

4.1. Primary Classification Task I - ADHD versus NC

No method performed best nor worst on all metrics as shown in Table 4. However, in three out of five submissions, we observed a significant drop in performance between validation and test data, whereas methods S2 and S5 were able to retain their performance. A drop when moving from seen to unseen data in the evaluated metrics is to be expected, however, considering the even split between patients and controls, many of the methods generalized to near-chance classification accuracy. Furthermore, the validation and test datasets were matched to the training data on all phenotypic variables. Hence, Table 4 suggests an overfitting to the released data.

Erring on the side of caution, frameworks intended for use in the clinic will often prefer to optimize on FDR (type I error) and FNR (type II error), instead of aiming to perform unilaterily well in all the presented metrics. If we were to assess performance by this criteria, Table 3 shows S5 outperforming other methods on the unseen ADHD test data (average of FDR and FNR: S1: 0.38; S2: 0.43; S3: 0.49; S4: 0.47; S5: 0.32). While S5 demonstrates the best average value, a FDR and FNR of 0.3 can still be seen as too high, where 30% of the patients and controls are misclassified.

We can further differentiate the methodologies based on the results of the calibration curves. Three out of five classifiers (S2, S3 and S4) produced a very narrow range of probabilities for classification (around 0.5 ± 0.1), whereas the two remaining classifiers utilized (almost) the full probability spectrum. Not utilizing the full predicted probability spectrum may reflect the models' uncertainty in classification and could potentially lead to misclassification in case of "noise" in the data. Thus, S5 demonstrated the overall best performance on unseen data. Furthermore, the calibration curve for S5 vaguely resembles a sigmoidal shape with an inflection point around a predicted probability of 0.6.

4.2. Transference Classification Task II - ASD with ADHD Comorbidity versus NC

Only one method (S5) exhibited the correct trend of predicted probabilities in classifying ASD patients with ADHD comorbidity (Figure 4B). All other methodologies predominantly assigned a higher probability of disease for NC, than for ASD patients. This may highlight the difficulties of achieving generalizability in these models if other comorbidities are present, and further agrees with our quantitative ranking of methodologies.

Overall, the large decrease in performances between Task I and Task II (Table 3), along with differences in the calibration curves (Figure 4) have two likely explanations. First, the models are overfitting to the released datasets. Second, the predictive rsfMRI features learned for ADHD do not transfer onto the ASD cohort with an ADHD co-morbidity, either because the latter has a unique neural phenotype, or because the commonalities between the disorders are overwhelmed by other signatures in the rsfMRI data, which did not allow the extraction of biologically meaningful information that could be specific to ADHD.

4.3. Approaches and Considerations for Future Challengers

While we cannot state unequivocally one submission that performed the best across all evaluation metrics in both Tasks (see Table 3), we can draw inferences from our submissions that may help to improve future developments of connectomic classification methodologies.

Connectivity matrix estimation varies across rsfMRI connectome studies and has been shown to affect classification performance (Abraham et al. (2017)). This lack of consensus in the field is reflected in our submissions - with each method using a different type of connectome as input. The

same observation can be made for choice of atlas, which can also influence performance (Abraham et al. (2017)). Interestingly, none of the CNI-TLC submissions used traditional graph metrics (Rubinov and Sporns (2010)), such as node degree, centrality, and small-worldness, which are ubiquitous in functional connectivity literature. Given that machine-/deep-learning methods can utilize big data during training, utilizing connectome data 'as is' seems intuitive, and it may be that traditional graph metrics are not favored for collapsing the number of data points in a time series. Even so, there may be advantages to including metrics that characterize complex global and topological properties unique to brain networks which may not be directly interpreted from raw connectivity matrices or timeseries data.

Despite the growing popularity of deep-learning methods, only one submission used an end-to-end deep-learning framework. Instead, the most common classifier among Challenge submissions was the support vector machine (Cortes and Vapnik (1995)). Ultimately, the deep-learning approach (S5) had the highest overall rank. Part of the reason may be attributed to an initial feature selection being performed and using only the most discriminative connections as input to the classifier. In addition, they employed a dynamic model that can capture key temporal information in the rsfMRI signal. The second place submission on classification Task I, in contrast to others, relied on the normalized graph Laplacian (S2). The graph Laplacian and its eigenspectrum enable a mapping of discrete data (a network) into vector spaces and manifolds, and their advantages have been greatly investigated, e.g., in fields such as clustering (Chung (1997)). This form of network representation has also found its way into connectomes and may have potential to capture further discriminative information (Abdelnour et al. (2014); Chung et al. (2016b,a); Schirmer and Chung (2019)). These data filtering or manipulation techniques are effective for their robustness to overfitting (Du et al. (2018)) and may be a reason for S5 and S2 achieving the best generalization performance from Task I to Task II.

We also observed that not all available information was utilized by participants. While the winning algorithm pre-selected the most discriminating features, this approach could be extended to feature selection across multiple atlases. This observation highlights the importance of including prior knowledge in developing classifiers for clinical tasks. Generally, we observed that most of the classifiers did not cover the full predicted probability spectrum. While this calibration may not be directly related to a classifier's performance on the actual task, it can serve as a proxy for assessing whether an

algorithm is extracting meaningful information from the data and can reveal issues in the classification process.

Another option for future challenges is the pooling of all classification results to asses their performance as an aggregate entity. As participants were asked to output classification scores, we combined all methodologies *post hoc* by obtaining their collective average, median, and maximum confidence (from the predicted value that is furthest away from 0.5), see Appendix C, Table C.5). Subsequently, if a consensus vote outperforms the individual methodologies, secondary analyses are required to fully understand the implications.

4.4. Challenge Contributions

The data used for this Challenge represents one of the largest collections of rsfMRI data acquired at a single site for three different cohorts (ADHD, ASD, and NC). The data was carefully and consistently collected on a single scanner model (Philips Achieva) by researchers at KKI. The in-house preprocessing pipeline is also standard in the field and included intermediate manual checks for data quality.

As part of our aim was to investigate the classification performance of several methodologies, we chose to release data that was uniformly processed by a single pipeline that has been previously published (see Muschelli et al. (2014); Nebel et al. (2016); Stoodley et al. (2017)). The effect of rsfMRI processing decisions on analysis is a large and important issue to address that is beyond the scope of our Challenge, but has been investigated by others (see e.g. Bowring et al. (2019); Výtvarová et al. (2017); Vergara et al. (2017)). We note that a subset of the rsfMRI data has been released through ABIDE (Di Martino et al. (2013)) and ADHD-200[9]. However, we were careful to ensure that the testing set contained only private data that had never been released and, as only pre-trained models were accepted, none of the algorithms were retrained with access to the test data. Each of our training, validation, and testing datasets were matched on demographic variables, in order to remove extraneous confounds. Finally, we opted to release the average time series from the rsfMRI data, rather than static or dynamic connectivity matrices, to provide more flexibility for participants. Likewise, we used three popular atlases to provide a range of spatial resolutions.

[9]http://fcon_1000.projects.nitrc.org/indi/adhd200/results.html

In order to standardize our evaluation, we implemented a novel framework for participants to submit pre-trained models as Docker images. The resulting Docker image from our framework also doubles as a plug-in compatible with the ChRIS platform[10], a pervasively open source framework that utilizes cloud technologies to democratize medical analytics application development. ChRIS allows researchers the ability to simply deploy the same application they have already developed in a cloud infrastructure with access to more data, more computational resources, and more collaboration to drive medical innovation, while standardizing healthcare application development. Subsequently, our approach gives participants the option to share their pre-trained model with the wider medical imaging research community.

One novel aspect of our Challenge, in comparison to current practices, is the large array of evaluation metrics used to assess the performance of each algorithm. This decision was motivated by a lack of consensus in the field about which metric is the "best" for evaluation, as it can be seen through the plethora of assessment metrics (Hossin and Sulaiman (2015)). Importantly, each metric assesses different aspects of an algorithm's performance. Another key strength of our evaluation procedure is that the test data is kept private. The unseen test data allows us to probe model overfitting and provides an objective comparison between algorithms.

With the above contributions, CNI-TLC has maintained and followed guidelines specifically outlined for challenges advocated by MICCAI with the intention of upholding transparency and fairness (Maier-Hein et al. (2018)).

4.5. Future Work

One of the main takeaways of our CNI-TL Challenge is that significant improvements are necessary in order to translate functional connectivity into clinical practice. Considering the performance across all participants summarized in Table 4, the overall scores are relatively low, with even the best results on the ASD cohort closely resembling chance in most metrics.

The CNI-TL Challenge highlights some important recommendations for future work on classification using functional connectivity. For example, our Challenge demonstrates the well-known issue of model generalizability to unseen data. This result supports the general recommendation that the data should be split into training, validation, and testing sets, with the testing

[10] https://chrisproject.org

set evaluated only once at the end of the study. Accordingly, we withheld the test set from CNI-TLC participants and continue to do so from the public, allowing the continuation of the Challenge, while methodologies can be further evaluated through the ChRIS platform.

By pooling submissions to solve the same problem, our Challenge has demonstrated the diversity and creativity of methods developed to use functional connectomes for classification. With specifics to model- and neural network-based methods, CNI-TLC further highlighted the need for a consensus on a series of experimental design tests to understand the influence of data representation (i.e. matrix estimation type, atlas, thresholding) in relation to an approach, disease, or dataset. Given the huge effort in recent years to amass and coalesce data from multiple sites, perhaps a similar ethos of combining and sharing efforts and approaches is equally necessary to tackle the neuroimaging challenges of today.

In recognition of the above points and of the role that challenges play in the field, we will keep the CNI-TL Challenge available and online for scientists to continually test their approaches. Submission is open to everyone, whether or not they have participated in the Challenge. Specifically, we provide a website[11] for developers of new methods to upload new solutions as ChRIS plug-ins, with an automated evaluation infrastructure "behind the scenes" offering dynamic feedback on the performance of their model on the test set. Importantly, the test data used is not exposed, thus providing a fair baseline from which to measure comparative performance.

5. Conclusion

The CNI-TL Challenge is an important step in the field of connectomics. First, it demonstrates the necessity of objective evaluation to assess generalizability to unseen data. Second, it goes beyond a single task-optimization setting to address the key question of whether we are capturing biologically meaningful phenomena. Here, we showed that the classification performance of all methods dropped nearly to chance on a patient population with the target disease as a comorbidity. This result underscores the need for further work to reach clinical translation of functional connectomics for disease identification. With the training and validation data remaining publicly

[11] https://fnndsc.childrens.harvard.edu/cnichallenge

available, and the test data accessible through an online evaluation platform, CNI-TLC facilitates continual development of new classification methods for connectomics.

Acknowledgements and Disclaimers

Organizers:

M. D. Schirmer was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 753896.

A. Venkataraman was supported by the National Science Foundation Collaborative Research in Computational Neuroscience (CRCNS) award 1822575 and the National Science Foundation CAREER award 1845430.

I. Rekik is supported by the 2232 International Fellowship for Outstanding Researchers Program of TUBITAK (Project No:118C288) and the the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101003403.

M. Kim was supported by UNC Greensboro New Faculty Award.

A. W. Chung was supported by the American Heart Association and Children's Heart Foundation Congenital Heart Defect Research Award, 19POST34380005.

Data team:

The patient recruitment, data acquisition, and preprocessing was supported by the Autism Speaks Foundation (awards 1739 and 2384) and by the National Institutes of Health under the following awards: K02 NS44850 (PI Mostofsky), R01 MH078160-10 (PI Mostofsky), R01 MH085328-09 (PI Mostofsky), R01 MH106564-03 (PI Edden), K23 MH101322-05 (PI Rosch), K23 MH107734-05 (PI Seymour), R01 NS048527-08 (PI Mostofsky), R01 NS096207-05 (PI Mostofsky), R01MH085328-14 (PI Mostofsky), U54HD079123, UL RR025005, and P54 EB15909.

This work was prepared while Karen Seymour was employed at Johns Hopkins University and Kennedy Krieger Institute. The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

Participants:

MeInternational: This work was supported by the EPSRC-funded UCL Centre for Doctoral Training in Medical Imaging (EP/L016478/1); the Na-

tional Institute for Health Research (NIHR); the Wellcome Trust (210182/Z/18/Z, 101957/Z/13/Z) and the Medical Research Council UK (Ref MR/J01107X/1).

HSE: This work was supported by the Russian Academic Excellence Project '5-100'.

ShefML: This work was supported by grants from the UK Engineering and Physical Sciences Research Council (EP/R014507/1).

YaleIPAG: N. C. Dvornek was supported by the National Institute of Health (NIH grants R01MH100028 and R01NS03519). J. Zhuang was supported by the National Institute of Health (NIH grant R01NS03519).

Appendix A.

Data used in this challenge were drawn from retrospective data acquired at the Kennedy Krieger Institute. The associated study names and IRB approval numbers are as follows:

1. Neurologic Basis of Inhibitory Deficits in ADHD (IRB 02-11-25-01)
2. Anomalous Motor Physiology in ADHD (IRB NA_00000292)
3. Deficient Response Control in ADHD (IRB NA_00027428)
4. Edden MRS (IRB NA_00088856)
5. Rosch Delay Discounting in ADHD (IRB 00032351)
6. Seymour Neurobehavioral Correlates of Frustration in ADHD (IRB 00063119)
7. Motor Skill Learning in Autism (IRB 03-05-27-10)
8. Motor Skill Learning in Autism: Assessment and Treatment of Altered Patterns of Learning (IRB NA_00027073)
9. Tactile Adaptation in Tourette's Syndrome (IRB NA_00090977)

Appendix B.

Detailed description of the methodology used by each submission.

Submission (S1) - MeInternational - Linear Support Vector Machine Framework to Predict Abnormal Functional Connectivity

Team: H. Irzan, M. Hütel, S. Ourselin, N. Marlow, A. Melbourne

The team evaluated different pipelines to select the best combination of functional connectivity metrics, atlases and classification algorithms based on

prediction accuracy. The nodes of the functional connectome are the individual atlas defined ROIs. The authors utilized the three different anatomical brain atlases (Craddock200 with 200 ROIs, Harvard Oxford with 110 ROIs, and AAL with 116 ROIs) and examined correlation, covariance, partial correlation, precision, and tangent embedding as metrics for functional connectivity. The team investigated l_1 and l_2 norm SVMs, linear regression with either l_1 or l_2 regularization, random forest, k-nearest neighbor, and naive Bayes classifiers as classification algorithms. Pipelines were built by using permutations of functional connectivity matrices, anatomical atlas, and classification algorithms as inAbraham et al. (2017). The performance of each pipeline was evaluated based on its prediction accuracy score on the validation set. Figure B.5 shows the main steps of the methodology. The best performance was achieved using the AAL atlas, correlation metric, and SVM with l_2 regularization and penalty parameter of the error term $C = 3^{-5}$, as highlighted in red in Figure B.5.

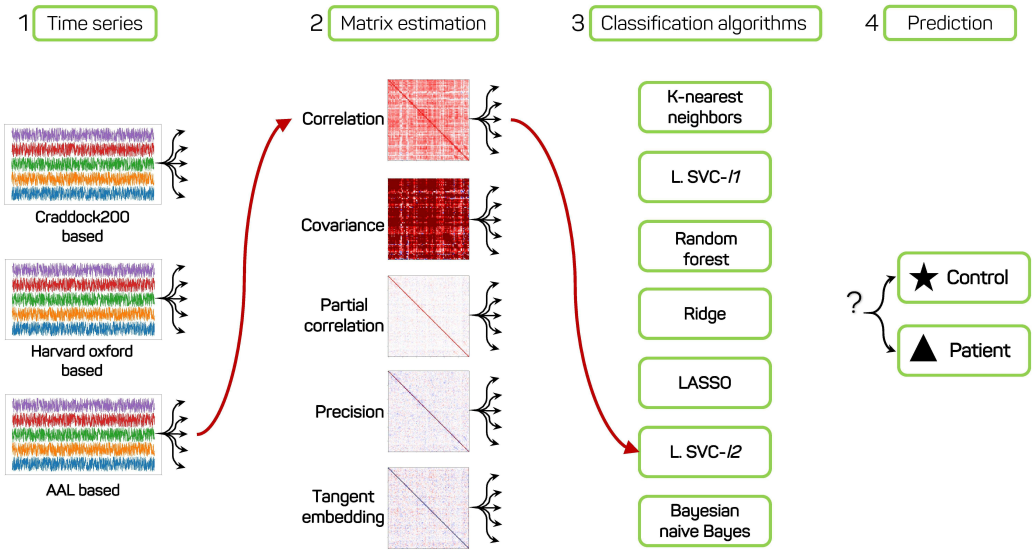

Figure B.5: Classification pipeline. In step 1, time series are produced by extracting the mean time courses of the preprocessed rs-fMRI using three anatomical atlases (Craddock200, Harvard Oxford, and AAL). These were provided as part of the challenge. In step 2, the time series are transformed into functional connectomes by using five matrix estimation methods. The functional connectomes are employed to perform supervised learning task with seven classifiers in step 3 and perform class prediction in step 4. The pipeline highlighted in red demonstrated the best classification accuracy on the test set.

Submission 2 (S2) - HSE - Classification of ADHD disorder Against Healthy Based on the Spectra of Normalized Laplacians

Team: E. Levchenko

The resting state time-varying signals of the rsfMRI data can be considered as a graph (connectome) (Sporns (2011)). Here, for each subject, a correlation matrix W_{ij} was calculated between each pair of ROIs i and j, using the Pearson correlation coefficient. First, the main diagonal in W and all negative values $W_{ij} < 0$ were set to zero, i.e. the graph was reduced to its positive weight subgraph (see e.g.Chung and Schirmer (2019)).

Graph theoretical studies have widely utilized the normalized Laplacian to characterize networks (see e.g.Chung and Graham (1997); Chung et al. (2016b)). The normalized Laplacian matrix L is defined as

$$L = D^{-1/2}(D - W)D^{-1/2}, \tag{B.1}$$

where each node in diagonal matrix D is given as $d_i = \sum_j A_{ij}$. The normalized Laplacian has several useful properties - one of them being that the eigenvalues are between 0 to 2 (Chung and Graham (1997)).

In this challenge, rsfMRI time series data were provided based on multiple atlases with different sizes for each subject. Here, the normalized Laplacian L was calculated for each subject and atlas independently. The distributions of eigenvalues of matrix L for each atlas were concatenated into one feature vector of length 426 for each subject. This feature vector serves as input for the classification problem, which was addressed using an SVM algorithm with a polynomial kernel. The hyperparameters were optimized on the provided training set and the classification accuracy was obtained on the validation set using 10-fold cross-validation.

Submission 3 (S3) - ShefML - Domain Independent SVM for CNI Challenge

Team: S. Zhou, M. Kunda, H. Lu

This solution learns a model via a two-stage pipeline: *feature extraction* and *classifier training*. For the feature extraction stage, Tangent Pearson (TP) (Kunda et al. (2020)) is applied to extract features from resting-state time-series of the AAL atlas. Specifically, as illustrated in Fig. B.6, Pearson correlation of the ROI-to-ROI relationship for each subject is computed first. Then tangent correlation analysis (Varoquaux et al. (2010a)) is performed to learn the group-level features of the brain network correlations, i.e. the

Figure B.6: Extracting Tangent Pearson connectivity features from resting-state fMRI time-series.

connectivity of connectives. Additionally, the mean and standard deviation of each time series are concatenated to the TP features.

For classification, it is assumed that the decisions (patient or control) made by a classifier should be independent to subjects' information, such as gender and age, and therefore Side Information Dependence Regularization (SIDeR) learning framework (Zhou et al. (2020)) is used to leverage the subjects' phenotype information for model training. The learning framework is given by

$$\min_{f} \underbrace{\mathcal{L}(f(\mathbf{X}^l), \mathbf{Y})}_{\text{Empirical risk}} + \underbrace{\sigma \|f\|_K^2}_{\text{Model complexity}} + \underbrace{\lambda \rho(f(\mathbf{X}), \mathbf{D})}_{\text{Side information dependence}}, \tag{B.2}$$

where $\rho(\cdot, \cdot)$ denotes a statistical independence metric, σ and λ are hyper-parameters, $\mathbf{X}^l$, $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{D}$ denote the labelled instances, training labels, all available instances, and side (phenotypic) information, respectively.

Three kinds of subject side information (gender, age, and handiness score) are selected and encoded as the matrix $\mathbf{D}$, where gender is encoded as 1 (male) and -1 (female), and the age and handedness scores are normalized to zero mean and unit variance. Empirically, Hinge (SVM) loss, ℓ_2 norm, and Hilbert-Schmidt Independence Criterion (HSIC;Gretton et al. (2005)) are employed for $\mathcal{L}(\cdot, \cdot)$, $\|f\|_K^2$, and $\rho(\cdot, \cdot)$, respectively. Therefore, the classification algorithm is a semi-supervised SVM trained on the labelled data, with the coefficients regularised by the HSIC between the classifier decision scores and subject phenotypic information.

Submission 4 (S4) - ShefML - Ensemble Model for CNI Challenge

Team: S. Zhou, M. Kunda, H. Lu

This approach contains three steps of analysis: *feature extraction, classifier training, and ensemble*. In the first step, five different type of features are extracted via computing the mean and standard deviation, Pearson correlation, tangent (Varoquaux et al. (2010a)), covariance (Varoquaux et al. (2010b)), and tangent Pearson correlation (Kunda et al. (2020)) from the fMRI time-series data of the AAL atlas. In the second step, five classifiers are trained on each type of feature extracted in step one respectively. The technical details here are exactly the same as in the classification method of S3 (see section Appendix B - Submission 3), which can be viewed as an alternative approach of this solution. In the last ensemble step, the final predictions are made by summarising the predictions given by the five classifiers from step two via majority voting. The pipeline is summarized in Figure B.7.

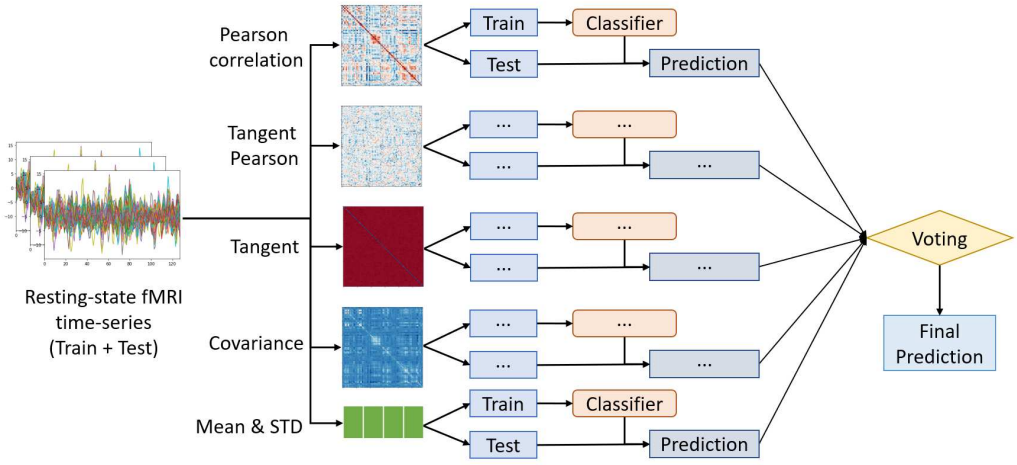

Figure B.7: The learning pipeline of Submission 4. Multiple features based on varying definitions of connectome creation were utilized in individual classifiers. Individual predictions were merged using majority voting.

Submission 5 (S5) - YaleIPAG - Learning Generalizable Recurrent Neural Networks from Small Task-fMRI Datasets

Team: N.C. Dvornek, J. Zhuang

The method (Fig. B.8) is based on Dvornek et al. (2018), which aims to learn generalizeable recurrent neural networks from small fMRI datasets. An LSTM-based network learns directly from the ROI time-series data, while

the demographic data is incorporated through subject-specific initialization of the LSTM hidden and cell states.

Using the AAL atlas, ROI selection was first performed by keeping ROIs whose connectivity (i.e., correlation) was consistently significantly different between ADHD and controls groups (two-sample t-test, $p < 0.05$) in 500 random subsamples of the data, using 90% of the training subjects in each subsample. Connectivity was considered consistently different if the p-value was in the top 2% of smallest p-values in each of the 500 subsamples. This process resulted in 22 ROIs.

The network architecture consisted of an LSTM layer with $M = 32$ hidden units and $T = 24$ timesteps (60s window), whose outputs were sent to a shared (across time) fully-connected layer with 1 node, followed by mean pooling and a sigmoid activation function to give the probability of ADHD. Time-series data from the 22 ROIs was used as input to the LSTM (Dvornek et al. (2017)), while demographic data was used for LSTM state initialization (Dvornek et al. (2018)). Specifically, demographic data was input to two fully-connected layers with M nodes each, representing the initial hidden and cell state of the LSTM.

To improve robustness, 10 models were trained using 10-fold cross-validation splits of the training dataset, with the validation dataset used to determine when to stop training. To predict on a new subject, all possible 60s windows of the time-series were input to the models to get a binary ADHD/control prediction for each window. The subject-level probability of ADHD for a single model was the proportion of windows labeled as ADHD for that model. Finally, the probability of ADHD for a given subject was computed as the mean of the 10 models' predictions.

Appendix C.

An example of pooling the performance from all Teams into an aggregate score by computing the average, median and maximum confidence of metrics across all classifiers. The maximum confidence was calculated by picking the classification where the predicted score is furthest away from 0.5, reflecting the confidence of an algorithm in its classification.

References

Abdelnour, F., Voss, H.U., Raj, A., 2014. Network diffusion accurately models the relationship between structural and functional brain connectivity

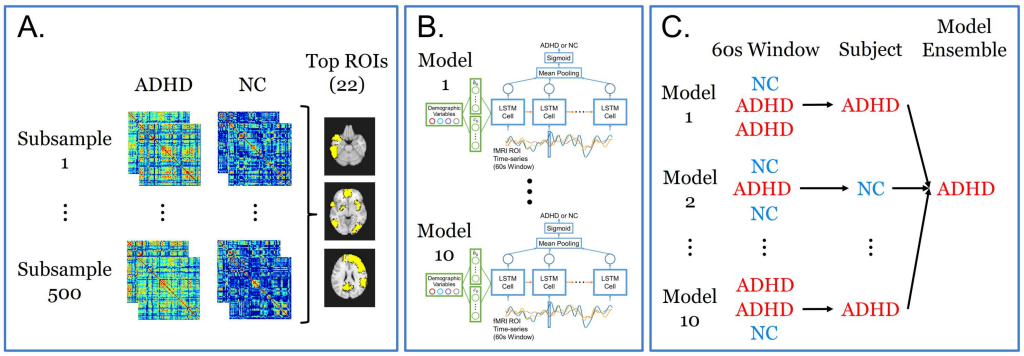

Figure B.8: Pipeline overview. A. Top AAL ROIs with pairwise correlations were selected which consistently differed between ADHD/NC using t-tests and repeated subject subsampling (90%). B. Ten LSTM models were trained using 10-fold split with top ROIs as input. C. Prediction of each subject using all 60s windows and model ensembles.

Table C.5: Consensus voting results of the challenge summarized as the median of each evaluation metric. All participants scores were combined using the mean (avg), median, and maximum confidence (maxconf, defined as classification based on the score furthest away from 0.5). In addition, the "Best median metric" out of all submissions from Table 4) is provided for comparison. The consensus vote does not outperform the best median metric in the ADHD cohort, while performing close to chance in the ASD classification task.

	ADHD				ASD			
	avg	maxconf	median	Best median metric	avg	maxconf	median	Best median metric
Acc	0.35	0.48	0.35	**0.66**	0.50	**0.55**	0.50	**0.55**
AUC	0.63	0.57	0.65	**0.68**	0.51	0.44	0.54	**0.56**
F1	0.32	0.48	0.32	**0.70**	0.51	0.53	0.52	**0.55**
FDR	0.67	0.52	0.67	**0.33**	0.50	**0.46**	0.50	**0.47**
FNR	0.70	0.50	0.65	**0.30**	0.50	0.50	0.45	**0.30**
FOR	0.65	0.52	0.65	**0.30**	0.50	**0.46**	0.50	0.47
FPR	0.65	0.50	0.65	**0.35**	0.50	**0.45**	0.55	**0.45**
GM	0.33	0.48	0.33	**0.70**	0.51	0.53	0.53	**0.56**
Inf.	-0.30	-0.05	-0.30	**0.35**	0.00	**0.10**	0.00	0.05
Mark	-0.30	-0.05	-0.30	**0.36**	0.00	**0.10**	0.00	0.05
MCC	-0.30	-0.05	-0.30	**0.36**	0.00	**0.10**	0.00	0.05
NPV	0.35	0.48	0.35	**0.70**	0.50	**0.54**	0.50	0.53
OP	13.86	18.95	13.86	**26.96**	20.00	**21.82**	19.90	20.86
Pre	0.33	0.48	0.33	**0.67**	0.50	**0.55**	0.50	0.53
Sen	0.30	0.50	0.30	**0.75**	0.50	0.50	0.55	**0.70**
Spec	0.35	0.50	0.35	**0.65**	0.50	**0.55**	0.45	**0.55**

networks. NeuroImage 90, 335–347. doi:10.1016/j.neuroimage.2013.12.039.

Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D.,

Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. NeuroImage 147, 736–745.

Americal Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (5th ed.). American Psychiatric Publishing, Arlington, VA.

Behzadi, Y., Restom, K., Liau, J., Liu, T., 2007. A component based noise correction method (compcor) for bold and perfusion based fMRI. NeuroImage 37, 90–101.

Bonkhoff, A.K., Schirmer, M.D., Bretzner, M., Etherton, M., Donahue, K., Tuozzo, C., Nardin, M., Giese, A.K., Wu, O., Calhoun, V., Grefkes, C., Rost, N.S., 2020. Dynamic functional connectivity analysis reveals transiently increased segregation in patients with severe stroke. medRxiv URL: https://www.medrxiv.org/content/early/2020/06/03/2020.06.01.20119263, doi:10.1101/2020.06.01.20119263.

Bowring, A., Maumet, C., Nichols, T.E., 2019. Exploring the impact of analysis software on task fmri results. Human brain mapping 40, 3362–3384.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., Cardoso, M.J., Cawley, N., Ciccarelli, O., Wheeler-Kingshott, C.A., Ourselin, S., Catanese, L., Deshpande, H., Maurel, P., Commowick, O., Barillot, C., Tomas-Fernandez, X., Warfield, S.K., Vaidya, S., Chunduru, A., Muthuganapathy, R., Krishnamurthi, G., Jesson, A., Arbel, T., Maier, O., Handels, H., Iheme, L.O., Unay, D., Jain, S., Sima, D.M., Smeets, D., Ghafoorian, M., Platel, B., Birenbaum, A., Greenspan, H., Bazin, P.L., Calabresi, P.A., Crainiceanu, C.M., Ellingsen, L.M., Reich, D.S., Prince, J.L., Pham, D.L., 2017. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. NeuroImage 148, 77 – 102. URL: http://www.sciencedirect.com/science/article/pii/S1053811916307819, doi:https://doi.org/10.1016/j.neuroimage.2016.12.064.

Chung, A.W., Pesce, E., Monti, R.P., Montana, G., 2016a. Classifying hcp task-fmri networks using heat kernels, in: 2016 International Workshop

on Pattern Recognition in Neuroimaging (PRNI), pp. 1–4. doi:10.1109/PRNI.2016.7552339.

Chung, A.W., Schirmer, M.D., 2019. Network dependency index stratified subnetwork analysis of functional connectomes: An application to autism, in: International Workshop on Connectomics in Neuroimaging, Springer. pp. 126–137.

Chung, A.W., Schirmer, M.D., Krishnan, M.L., Ball, G., Aljabar, P., Edwards, A.D., Montana, G., 2016b. Characterising brain network topologies: A dynamic analysis approach using heat kernels. NeuroImage 141, 490 – 501. doi:https://doi.org/10.1016/j.neuroimage.2016.07.006.

Chung, F.R., Graham, F.C., 1997. Spectral graph theory. 92, American Mathematical Soc.

Chung, F.R.K., 1997. Spectral Graph Theory. Conference Board of the Mathematical Society (CBMS), Number 92, American Mathematical Society.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Ameli, R., Ferré, J.C., et al., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Scientific reports 8, 13650.

Conners, C., 2002. Conners' Rating Scale Revised. Multi-Health Systems, Inc., Toronto, Canada.

Conners, C., 2008. Conners' 3. Multi-Health Systems, Inc., North Tonawanda, NY.

Cortes, C., Vapnik, V., 1995. Support vector networks. Machine Learning 20, 273–297.

Cox, R.W., 1996. Afni: Software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research 29, 162 – 173.

Craddock, R.C., James, G.A., Holtzheimer III, P.E., Hu, X.P., Mayberg, H.S., 2012. A whole brain fmri atlas generated via spatially constrained spectral clustering. Human brain mapping 33, 1914–1928.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31, 968–980.

Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F., Alaerts, K., Anderson, J., Assaf, M., Bookheimer, S., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D., Gallagher, L., Kennedy, D., Keown, C., Keysers, C., Lainhart, J., Lord, C., Luna, B., Menon, V., Minshew, N., Monk, C., Mueller, S., Muller, R.A., Nebel, M., Nigg, J., O'Hearn, K., Pelphrey, K., Peltier, S., Rudie, J., Sunaert, S., Thioux, M., Tyszka, J., Uddin, L., Verhoeven, J., Wenderoth, N., Wiggins, J., Mostofsky, S., Milham, M., 2013. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry epub.

Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry 19, 659.

Douw, L., van Dellen, E., Gouw, A.A., Griffa, A., de Haan, W., van den Heuvel, M., Hillebrand, A., Van Mieghem, P., Nissen, I.A., Otte, W.M., Reijmer, Y.D., Schoonheim, M.M., Senden, M., van Straaten, E.C.W., Tijms, B.M., Tewarie, P., Stam, C.J., 2019. The road ahead in clinical network neuroscience. Network Neuroscience 3, 969–993. doi:10.1162/netn_a_00103.

Dowell, L., Mahone, M., Mostofsky, S., 2009. Associations of postural knowledge and basic motor skill with dyspraxia in autism: Implication for abnormalities in distributed connectivity and motor learning. Neuropsychology 23, 563–570.

D'Souza, N., Nebel, M., Wymbs, N., Mostofsky, S., Venkataraman, A., 2019. Integrating neural networks and dictionary learning for multidimensional clinical characterizations from functional connectomics data, in: MICCAI: Medical Image Computing and Computer Assisted Intervention, p. Accepted.

Du, Y., Fu, Z., Calhoun, V.D., 2018. Classification and prediction of brain disorders using functional connectivity: Promising but challenging. Frontiers in Neuroscience 12, 525. doi:10.3389/fnins.2018.00525.

DuPaul, G., Power, T., Anastopoulos, A., Reid, R., 1998. ADHD Rating Scale IV: Checklists, Norms, and Clinical Interpretation. Guilford Press.

Dvornek, N.C., Ventola, P., Pelphrey, K.A., Duncan, J.S., 2017. Identifying autism from resting-state fmri using long short-term memory networks, in: MLMI 2017, pp. 362–370.

Dvornek, N.C., Yang, D., Ventola, P., Duncan, J.S., 2018. Learning generalizable recurrent neural networks from small task-fmri datasets, in: MICCAI 2018, pp. 329–337.

D'Souza, N., Nebel, M., Wymbs, N., Mostofsky, S., Venkataraman, A., 2020. A joint network optimization framework to predict clinical severity from resting state functional MRI data. NeuroImage 206, 116314.

Essen, D.V., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S., Penna, S.D., Feinberg, D., Glasser, M., Harel, N., Heath, A., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S., Prior, F., Schlaggar, B., Smith, S., Snyder, A., Xu, J., Yacoub, E., 2012. The human connectome project: A data acquisition perspective. NeuroImage 62, 2222 – 2231.

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W. (Eds.), 2007. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press. URL: http://store.elsevier.com/product.jsp?isbn=9780123725608.

Gotham, K., Risi, S., Pickles, A., Lord, C., 2007. The autism diagnostic observation schedule: Revised algorithms for improved diagnostic validity. J Autism and Developmental Disorders 37, 613–627.

Gretton, A., Bousquet, O., Smola, A., Schölkopf, B., 2005. Measuring statistical dependence with Hilbert-Schmidt norms, in: International Conference on Algorithmic Learning Theory, Springer. pp. 63–77.

Hamed, A.M., Kauer, A.J., Stevens, H.E., 2015. Why the diagnosis of attention deficit hyperactivity disorder matters. Frontiers in psychiatry 6, 168.

Heimann, T., Van Ginneken, B., Styner, M.A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., et al., 2009. Comparison and evaluation of methods for liver segmentation from ct datasets. IEEE transactions on medical imaging 28, 1251–1265.

Hossin, M., Sulaiman, M., 2015. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process 5, 1.

Huerta, M., Lord, C., 2012. Diagnostic evaluation of autism spectrum disorders. Pediatric Clinics of North America 59, 103.

Kaiser, M., 2017. Mechanisms of connectome development. Trends in Cognitive Sciences 21, 703–717. URL: https://doi.org/10.1016/j.tics.2017.05.010, doi:10.1016/j.tics.2017.05.010.

Kaufman, J., Birmaher, B., Brent, D., Ryan, N., Rao, U., 2000. K-sads-pl. Journal of the American Academy of Child and Adolescent Psychiatry 39, 1208.

Ktena, S.I., Schirmer, M.D., Etherton, M.R., Giese, A.K., Tuozzo, C., Mills, B.B., Rueckert, D., Wu, O., Rost, N.S., 2019. Brain connectivity measures improve modeling of functional outcome after acute ischemic stroke. Stroke 50, 2761–2767.

Kunda, M., Zhou, S., Gong, G., Lu, H., 2020. Improving multi-site autism classification based on site-dependence minimisation and second-order functional connectivity. bioRxiv .

Lake, E.M., Finn, E.S., Noble, S.M., Vanderwal, T., Shen, X., Rosenberg, M.D., Spann, M.N., Chun, M.M., Scheinost, D., Constable, R.T., 2019. The functional brain organization of an individual allows prediction of measures of social abilities transdiagnostically in autism and attention-deficit/hyperactivity disorder. Biological Psychiatry 86, 315 – 326. doi:https://doi.org/10.1016/j.biopsych.2019.02.019. autism Spectrum Disorder: Mechanisms and Features.

Leitner, Y., 2014. The co-occurrence of autism and attention deficit hyperactivity disorder in children–what do we know? Frontiers in human neuroscience 8, 268.

Lord, A., Horn, D., Breakspear, M., Walter, M., 2012. Changes in community structure of resting state functional connectivity in unipolar depression. PLoS One 7, e41282.

Lord, C., Rutter, M., Couteur, A., 1994. Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism and Developmental Disorders 24, 659–685.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., Feldmann, C., Frangi, A.F., Full, P.M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B.A., März, K., Maier, O., Maier-Hein, K., Menze, B.H., Müller, H., Neher, P.F., Niessen, W., Rajpoot, N., Sharp, G.C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A.A., van der Sommen, F., Wang, C.W., Weber, M.A., Zheng, G., Jannin, P., Kopp-Schneider, A., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nature Communications 9, 1. doi:10.1038/s41467-018-07619-7.

Martino, A.D., Zuo, X.N., Kelly, C., Grzadzinski, R., Mennes, M., Schvarcz, A., Rodman, J., Lord, C., Castellanos, F.X., Milham, M.P., 2013. Shared and distinct intrinsic functional network centrality in autism and attention-deficit/hyperactivity disorder. Biological Psychiatry 74, 623 – 632. doi:https://doi.org/10.1016/j.biopsych.2013.02.011.

McPartland, J., Wu, J., Bailey, C., Mayes, L., Schultz, R., Klin, A., 2011. Atypical neural specialization for social percepts in autism spectrum disorder. Social Neuroscience Epub, 1–16.

Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Breeuwer, M., Bouvy, W.H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A., et al., 2015. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. Computational intelligence and neuroscience 2015, 1.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging 34, 1993–2024.

Mill, R.D., Ito, T., Cole, M.W., 2017. From connectome to cognition: The search for mechanism in human functional brain networks. NeuroImage 160, 124 – 139. URL: http://www.sciencedirect.com/science/article/pii/S1053811917300836, doi:https://doi.org/10.1016/j.neuroimage.2017.01.060. functional Architecture of the Brain.

Murphy, K., Van Ginneken, B., Reinhardt, J.M., Kabus, S., Ding, K., Deng, X., Cao, K., Du, K., Christensen, G.E., Garcia, V., et al., 2011. Evaluation of registration methods on thoracic ct: the empire10 challenge. IEEE transactions on medical imaging 30, 1901–1920.

Muschelli, J., Nebel, M.B., Caffo, B.S., Barber, A.D., Pekar, J.J., Mostofsky, S.H., 2014. Reduction of motion-related artifacts in resting state fmri using acompcor. NeuroImage 96, 22 – 35. URL: http://www.sciencedirect.com/science/article/pii/S105381191400175X, doi:https://doi.org/10.1016/j.neuroimage.2014.03.028.

Nebel, M.B., Eloyan, A., Nettles, C.A., Sweeney, K.L., Ament, K., Ward, R.E., Choe, A.S., Barber, A.D., Pekar, J.J., Mostofsky, S.H., 2016. Intrinsic visual-motor synchrony correlates with social deficits in autism. Biological Psychiatry 79, 633 – 641. URL: http://www.sciencedirect.com/science/article/pii/S0006322315007283, doi:https://doi.org/10.1016/j.biopsych.2015.08.029. cortical Function and Social Deficits in Autism.

Oldfield, R., 1971. The assessment and analysis of handedness: The edinburgh inventory. Neuropschologia 9, 97–113.

Pelphrey, K., Yang, D.J., McPartland, J., 2014. Building a social neuroscience of autism spectrum disorder. Current Topics in Behavioral Neuroscience 16, 215–233.

Pervaiz, U., Vidaurre, D., Woolrich, M.W., Smith, S.M., 2019. Optimising network modelling methods for fmri. bioRxiv , 741595.

Ray, S., Miller, M., Karalunas, S., Robertson, C., Grayson, D.S., Cary, R.P., Hawkey, E., Painter, J.G., Kriz, D., Fombonne, E., Nigg, J.T., Fair, D.A., 2014. Structural and functional connectivity of the human brain in autism spectrum disorders and attention-deficit/hyperactivity disorder: A rich club-organization study. Human Brain Mapping 35, 6032–6048. doi:10.1002/hbm.22603, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.22603.

Reich, W., Welner, Z., Herjanic, B., 1997. Diagnostic Interview for Children and Adolescents-IV. Springer, North Towaganda Falls, NY.

Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: Uses and interpretations. NeuroImage 52, 1059–1069.

Schirmer, M.D., Chung, A.W., 2019. Heat kernels with functional connectomes reveal atypical energy transport in peripheral subnetworks in autism, in: Schirmer, M.D., Venkataraman, A., Rekik, I., Kim, M., Chung, A.W. (Eds.), Connectomics in NeuroImaging, Springer International Publishing, Cham. pp. 54–63.

Schirmer, M.D., Ktena, S.I., Nardin, M.J., Donahue, K.L., Giese, A.K., Etherton, M.R., Wu, O., Rost, N.S., 2019. Rich-club organization: An important determinant of functional outcome after acute ischemic stroke. Frontiers in Neurology 10, 956. doi:10.3389/fneur.2019.00956.

Shine, J., Bissett, P., Bell, P., Koyejo, O., Balsters, J., Gorgolewski, K., Moodie, C., Poldrack, R., 2016. The dynamics of functional brain networks: Integrated network states during cognitive task performance. Neuron 92, 544–554. URL: https://doi.org/10.1016/j.neuron.2016.09.018, doi:10.1016/j.neuron.2016.09.018.

Somerville, L.H., Bookheimer, S.Y., Buckner, R.L., Burgess, G.C., Curtiss, S.W., Dapretto, M., Elam, J.S., Gaffrey, M.S., Harms, M.P., Hodge, C., Kandala, S., Kastman, E.K., Nichols, T.E., Schlaggar, B.L., Smith, S.M., Thomas, K.M., Yacoub, E., Essen, D.C.V., Barch, D.M., 2018. The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. NeuroImage 183, 456 – 468. URL: http://www.sciencedirect.com/science/article/pii/S1053811918307481, doi:https://doi.org/10.1016/j.neuroimage.2018.08.050.

Sporns, O., 2011. The human connectome: a complex network. Annals of the New York Academy of Sciences 1224, 109–125.

Stoodley, C.J., D'Mello, A.M., Ellegood, J., Jakkamsetti, V., Liu, P., Nebel, M.B., Gibson, J.M., Kelly, E., Meng, F., Cano, C.A., et al., 2017. Altered cerebellar connectivity in autism and cerebellar-mediated rescue of autism-related behaviors in mice. Nature neuroscience 20, 1744–1751.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. Neuroimage 15, 273–289.

Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P., Thirion, B., 2010a. Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 200–208.

Varoquaux, G., Gramfort, A., Poline, J.B., Thirion, B., 2010b. Brain covariance selection: better individual functional connectivity models using population prior, in: Advances in neural information processing systems, pp. 2334–2342.

Venkataraman, A., Duncan, J., Yang, D., Pelphrey, K., 2013a. An unbiased bayesian approach to functional connectomics implicates social-communication networks in autism. NeuroImage Clinical 8, 356–366.

Venkataraman, A., Kubicki, M., Golland, P., 2013b. From brain connectivity models to region labels: Identifying foci of a neurological disorder. IEEE Transactions on Medical Imaging 32, 2078–2098.

Venkataraman, A., Rathi, Y., Kubicki, M., Westin, C.F., Golland, P., 2012. Joint modeling of anatomical and functional connectivity for population studies. IEEE Transactions on Medical Imaging 31, 164–182.

Venkataraman, A., Yang, D., Pelphrey, K., Duncan, J., 2016. Bayesian community detection in the space of group-level functional differences. IEEE Transactions on Medical Imaging 35, 1866–1882.

Vergara, V.M., Mayer, A.R., Damaraju, E., Hutchison, K., Calhoun, V.D., 2017. The effect of preprocessing pipelines in subject classification and detection of abnormal resting state functional network connectivity using group ica. NeuroImage 145, 365–376.

Výtvarová, E., Fousek, J., Bartoň, M., Mareček, R., Gajdoš, M., Lamoš, M., Nováková, M., Slavíček, T., Peterlik, I., Mikl, M., 2017. The impact of diverse preprocessing pipelines on brain functional connectivity, in: 2017 25th European Signal Processing Conference (EUSIPCO), pp. 2644–2648.

Wechsler, D., 2003. The Wechsler Intelligence Scale for Children—Fourth Edition. Pearson, London, UK.

Wolterink, J.M., Leiner, T., De Vos, B.D., Coatrieux, J.L., Kelm, B.M., Kondo, S., Salgado, R.A., Shahzad, R., Shu, H., Snoeren, M., et al., 2016. An evaluation of automatic coronary artery calcium scoring methods with cardiac ct using the orcascore framework. Medical physics 43, 2361–2373.

Zhou, S., Li, W., Cox, C.R., Lu, H., 2020. Side information dependence as a regularizer for analyzing human brain conditions across cognitive experiments, in: Thirty-Forth AAAI Conference on Artificial Intelligence, pp. 6957–6964.