Robust, generalizable, and interpretable AI-derived brain fingerprints of autism and social-communication symptom severity

Supplemental Information

Contents

Supplementary Methods

ABIDE We leveraged neuroimaging and phenotypic data from the Autism Brain Imaging Data Exchange (ABIDE; http://fcon_1000.projects.nitrc.org/indi/abide/) (1, 2). The subject selection procedure is illustrated in **Supplementary Figure S1**. 419 subjects with ASD and 415 TD subjects from 14 sites were selected (Number of subjects per site: Mean = 60, Range: 26 to 180). **Supplementary Table S2** shows demographic information.

Stanford An independent cohort of participants recruited and scanned at Stanford University was used to investigate the generalizability of stDNN to previously unseen data. ASD diagnosis was assessed using the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R) as described in our published studies (3-5). 270 ASD subjects with resting-state fMRI data were identified, from which we chose 101 unique subjects with age and gender distributions, as well as head movement during fMRI, matched to ASD participants from the ABIDE cohort. The group matching was performed using an in-house matching program that uses the principles of genetic algorithms . 423 TD subjects with resting-state fMRI data were identified, from which we chose 101 unique subjects such that age and gender distributions, as well as head movement during fMRI were matched to the TD subjects from the ABIDE cohort. The subject selection procedure is illustrated in **Supplementary Figure S2**. **Supplementary Table S2** shows demographic information.

GENDAAR An independent cohort of participants acquired by the GENDAAR consortium and shared through the National Institutes of Mental Health Data Archive was used to further investigate the generalizability of stDNN to previously unseen data. ASD diagnosis was

confirmed using the ADOS and the ADI-R. 87 ASD subjects with resting-state fMRI data were identified, from which we chose 50 unique subjects with age and gender distributions as well as head movement during fMRI, matched to ASD participants from the ABIDE cohort using the in-house matching program (6). 70 TD subjects with resting-state fMRI data were identified, from which we chose 40 unique subjects such that age and gender distributions as well as head movement during fMRI were matched to the TD subjects from the ABIDE cohort. The subject selection procedure is illustrated in **Supplementary Figure S3**. **Supplementary Table S2** shows demographic information.

ADHD200-NYU. An independent cohort of participants recruited and scanned at New York University (NYU) and made available through the ADHD200 consortium (http://fcon_1000.projects.nitrc.org/indi/adhd200/) was used to investigate the specificity of the stDNN. Out of the eight independent sites who participated in the ADHD-200 collection, only the NYU site (i) had data acquisition parameters that matched the ABIDE and Stanford cohorts, (ii) included a large number of subjects (N > 200 subjects), (iii) had ASD and TD participants that were part of the ABIDE cohort, and was therefore used in this study. The Schedule of Affective Disorders and Schizophrenia for Children–Present and Lifetime Version (KSADS-PL) was used for ADHD diagnosis. **Supplementary Table S22** shows demographic information.

fMRI preprocessing

All functional MRI data were preprocessed by using SPM12 software package, as well as in-house MATLAB scripts. Structural MRI images were segmented into grey matter, white matter (WM), and cerebrospinal fluid (CSF). Prior to preprocessing, QA of functional and structural MRI was performed and subjects with poor quality imaging data were excluded from analysis. Resting-state functional MRI (fMRI) data were realigned to the averaged time frame to correct for head motion, slice-time corrected to the first slice, and co-registered to each participant's T1-weighted images. The functional images were then normalized to the standard Montreal Neurological Institute (MNI152) template at $2mm^3$. A 6-mm Gaussian kernel was used to spatially smooth the functional images and a band-pass filter ranging from 0.01 to 0.1 Hz was applied. Band-pass filtering of fMRI timeseries was used to remove low frequency artifacts such as scanner drifts and high frequency components, which do not contain useful information. Critically, band-pass filtering does not remove non-stationarities in the data, and non-stationarities such as time-varying means and covariances can still exist in a band pass filtered signal. To account for artifacts from motion and nonneural sources, the mean time series from each of the CSF and WM masks as well as 6 motion parameters, obtained by rigid body registration, were regressed out from the fMRI data. We used the binarized WM and CSF tissue probability maps provided by FSL.

Data input into the stDNN

We used the Brainnetome Atlas (246 regions) and computed the average resting-state fMRI timeseries across the voxels in a given region of interest (ROI). We used Brainnetome as it provides fine-grained brain-wide parcellations of both cortical and subcortical areas with better anatomical and functional interpretability than most other atlases. Critically, the Brainnetome

Atlas is one of the most extensively used atlases, with over 1000 studies using it (7), enabling the comparison of our method/findings with those from extant related research work as well as those under development elsewhere. Each participant's time series data was represented by a matrix of size $N_C \times N_T$, where N_C is the number of channels or ROIs and N_T is the number of time points. For participants belonging to the training (ABIDE) cohort, N_T was fixed to 180 (i.e. 360 seconds) to reduce computational complexity. To demonstrate that our stDNN model could accommodate varying time lengths, subjects in the testing cohorts were not subjected to this restriction. We used site information as well as sex and age as covariates. To demonstrate that our stDNN findings are robust to atlas selection, we also examined three commonly used atlases (covering a broad range of number of ROIs), including Schaeffer (1000 regions), CC200 (200 regions) and CC400 (392 regions).

stDNN model

We developed an innovative stDNN model to extract informative brain dynamics features that accurately distinguish between ASD and neurotypical controls. A key advantage of our approach is that it provides a novel technique to capture latent dynamics without the need for explicit feature engineering (8). Our stDNN model consists of two 1D convolutional blocks, a "temporal averaging" operation, and then a sigmoid layer for binary classification (**Figure 2**). Each convolutional block layer consists of a convolutional operation and ReLU activation.
We introduce "maxpool" layers after each of the two convolutional block layers. The input to the stDNN is each subject's $N_C \times N_T$ ROI timeseries matrix where $N_C = 246$ for Brainnetome Atlas. The first 1D CNN layer transforms the input spatio-temporally to $256 \times N_T$ with a

temporal kernel of size $F = 5$ and with 256 number of filters. In other words, the input is convolved with a filter of size 256×5 and produces an output of $256 \times \mathrm{N_T}$. The convolution operation is mathematically defined as:

$$\tilde{\boldsymbol{x}}_{l+1}(n) = \sum_{k=0}^{K} h_l(k)\boldsymbol{x}_l(n-k)$$

Where, the output of the convolution $\tilde{\boldsymbol{x}}_{l+1}$ at the l-th layer is defined as a linear combination of the convolutional kernel weights $h_l(k)'s$ and the output of the previous maxpool layer $\boldsymbol{x}_l$. This convolution operation projects the input data into multiple frequency bands. This 1D convolution layer is followed by a ReLU nonlinear operator defined as below:

$$\tilde{\boldsymbol{x}}_{l+1}(n) - \max(0, \tilde{\boldsymbol{x}}_{l+1}(n))$$

which results in an output of size $256 \times \mathrm{N_T}$. This nonlinear operation helps in extracting nonlinear features in the data. In modern deep learning architectures, ReLU is preferred over other nonlinear operators such as sigmoid and tanh functions. This is because ReLU, as opposed to the sigmoid and tanh functions, does not saturate the gradients and therefore the backpropagation algorithm can effectively learn the model weights. The ReLU nonlinear operation is followed by a maxpool layer with a kernel size of 2, which produces an output data of size $256 \times \mathrm{N_T}$. The "maxpool" layers help in (a) reducing the temporal dimension of the data, (b) hierarchical representation of the features, and (c) increasing the receptive field of the filter to capture the long-term correlations in the timeseries. The output of this maxpool layer is an input to the second 1D convolution block. The processing of this block is similar to the first block with a difference that the number of filters used in the 1D CNN layer is 512 with a temporal kernel size of 3. The output of this block after ReLU and the maxpool layer is $512 \times \mathrm{N_T}$. The output of the second maxpool layer is to a "temporal averaging layer". Conventionally, after the last

convolutional block, the data is flattened and a fully connected layer is connected to the output sigmoid layer. The fully connected layers typically have the maximum number of parameters to be trained compared to the convolutional layers. In our model, instead of the normal flattening operation, we use a "temporal averaging layer" where we average the temporal features for each filter and therefore the number of inputs to the fully connected layer is just the number of output channels of the second convolution block layer. The advantages of averaging layer over the flattening layer are (a) the number of parameters reduced from $N_{C2} \times N_{T2}$ to N_{C2} where N_{C2} is the number of output channels from the second convolutional block layer which is 512 and N_{T2} is the temporal dimension of the output of the second "maxpool" layer, (b) with averaging layer, we can train and test fMRI timeseries with varying time lengths. Temporal averaging layer is a dimensionality reduction step in the latent space and not in the original timeseries space, so is unlikely to cause loss of significant temporal information. Varying time length is common with open-source data where the data is acquired with different data acquisition protocols. We introduce a dropout layer (= 0.5) before the fully connected layer to avoid overfitting during the model training process. To account for site-related heterogeneity, site information encoded with a one-hot encoding scheme was given as an input to the final fully connected layer. To account for age- and sex-related heterogeneity, sex information encoded with a one-hot encoding scheme and age were also given as inputs to the final fully connected layer. stDNN classified participants in the two groups by minimizing the binary cross-entropy cost function. We trained the model for up to 300 epochs with a stopping criterion and a learning rate of 0.00005 with a batch size of 32. An Adam optimizer with a weight decay of 0.0001 was used to estimate the stDNN model parameters (9).

Five-fold cross-validation classification analysis of ABIDE cohort data

To prevent bias and account for low variance, we conducted a five-fold cross-validation to evaluate the performance of our stDNN model. In the five-fold cross-validation approach, we divided the whole dataset into five different parts. Since we used a one-hot-encoding approach to address site and sex-related heterogeneity, neither site nor sex were stratified in the five parts. We used four parts for training and validation and the fifth part as the test set. We then rotate through the whole dataset five times to select a different section as the test set during each iteration (**Figure 3).** For each of the five subsets, we evaluate the performance of our stDNN model individually and then average over the five subsets to report the mean and standard deviation values of the key performance metrics (accuracy, precision, recall, F1). Using the five-fold cross-validation approach, the performance for every sample from the ABIDE data gets accounted, which helps in assessing the effectiveness of the model more robustly instead of just reporting the performance on one-time random split of the data **(Supplementary Table 1)**.

Classification analysis of Stanford cohort data using five-fold ABIDE cohort models

Similar to the five-fold cross-validation process used for ABIDE, for reporting the performance of our stDNN for Stanford cohort, we used each of the five stDNN models trained on different subsets of ABIDE. Using the five different models, we evaluate each model's performance on the Stanford cohort data independently (**Figure 3**) and report the mean and standard deviation values of the key performance metrics (accuracy, precision, recall, F1).

Classification analysis of GENDAAR cohort data using five-fold ABIDE cohort models

Similar to the five-fold cross-validation process used for ABIDE, for reporting the performance of our stDNN for GENDAAR cohort, we used each of the five stDNN models trained on different subsets of ABIDE. Using the five different models, we evaluate each model's performance on the GENDAAR cohort data independently (**Figure 3**) and report the mean and standard deviation values of the key performance metrics (accuracy, precision, recall, F1).

Classification analysis of ADHD200-NYU cohort data using five-fold ABIDE cohort models

Similar to the five-fold cross-validation process used for ABIDE, for reporting the performance of our stDNN for ADHD200-NYU cohort, we used each of the five stDNN models trained on different subsets of ABIDE. Using the five different models, we evaluate each model's performance on the ADHD200-NYU cohort data independently and report the mean and standard deviation values of the key performance metrics (accuracy, precision, recall, F1).

Control analysis examining alternative classification models

stDNN models the dynamic spatiotemporal characteristics of brain activity to classify ASD vs. neurotypical controls using fMRI timeseries without any explicit feature engineering. To demonstrate the advantages of our stDNN model over extant classification approaches, we performed extensive control analyses. We evaluated several commonly used linear and nonlinear classification algorithms, including K-Nearest Neighbor, Decision Tree, Linear SVM, Logistic Regression, Ridge Classifier, LASSO, and Random Forest. fMRI timeseries features were

provided as input to these classification algorithms. We explored several commonly used fMRI timeseries feature spaces, including static functional connectivity, amplitude of low-frequency fluctuation (ALFF), BOLD signal variability, and sliding-window functional connectivity variability as well as the raw fMRI timeseries. Static functional connectivity (10), which represents time-invariant functional interactions between brain regions, was computed by Pearson correlations between regional fMRI timeseries. ALFF (11), which reflects the intensity of spontaneous fluctuations in a brain region, was computed by obtaining the square root of the fMRI timeseries signal across the low-frequency range of 0.01-0.08 Hz for each of the 246 brainnetome brain regions. BOLD signal variability (12), which represents the moment-to-moment brain variability, was calculated as the standard deviation of the regional fMRI signal across the duration of the fMRI scan. Sliding-window functional connectivity variability, which reflects the variability in dynamic functional interactions between brain regions, was computed by first calculating time-varying functional connectivity using a sliding-window approach (13) and then calculating the standard deviation of the time-varying functional connectivity. To evaluate the performance of the aforementioned classification models in distinguishing between ASD and TD controls, we used the same five-fold cross-validation approach we used to evaluate our stDNN model. Briefly, we divided the whole dataset into five different parts, where we use four parts for training and validation and the fifth part as the test set. We then rotate through the whole dataset five times to select a different section as the test set during each iteration**.** For each of the five subsets, we evaluate the performance of the alternate classification model individually and then average over the five subsets to report the mean and standard deviation values of the key performance metrics (accuracy, precision, recall, F1).

Control analysis examining the relationship between head movement and classification

To demonstrate that our classification results are robust to movement confounds, we compared head movement, as measured by mean framewise displacement (FD), between participants who were correctly classified by our stDNN model and those who were incorrectly classified.

Identifying brain features underlying ASD classification

We used an integrated gradients (IG)-based feature attribution approach (14-18) to identify brain features that discriminated between the ASD and TD groups. A major problem in developing and evaluating feature attribution methods is that it is difficult to distinguish errors from the DNN model and those from feature attribution procedures. IG solves this problem by taking an approach that satisfies two fundamental axioms – sensitivity and implementation invariance (14-18). Another advantage of IG is that the gradients can be computed easily for any given network architecture. IG estimates the integral of gradients with respect to the *i-th* dimension of the input *x* along the straight-line path from a given (or random) baseline to the input as follows:

$$IG_i = (x - x') \int_0^1 (x_i - x'_i) \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \partial\alpha$$

where, IG_i is the integrated gradient for the *i-th* component of the input *x* and x' is the baseline input for which the neural network *F* results in a neutral output. IG provides a score of how important each feature contributes to the final prediction. This approach provides insights about

important features that predict ASD class label. Conventional gradient-based approaches wrongly assign zero attributions for inputs where the function is flat, even when the output of *F* for such an input is different from the baseline. IG avoids this problem by computing an average gradient along a linear path. Our IG implementation is based on the *"Captum" (https://captum.ai/docs/introduction.html)* module of *Pytorch.* The IG-derived feature importance/weights are computed at an individual level and relative to a baseline that is common across individuals, and therefore were not normalized.

Control analysis examining the relationship between head movement and brain features underlying ASD classification

We computed the squared distance correlation ($dcor^2$) (19) between the strength of features and the mean framewise displacement (FD). Specifically, $dcor^2$ between the strength of the top 5% of features and the mean FD was computed in the ASD group for the two cohorts separately. Briefly, $dcor^2$ is a measure of the nonlinear relationship between multidimensional variables, making it a better measure than conventional metrics like Pearson correlation, which only capture univariate linear relationships. $dcor^2$ has a range from 0 to 1, with $dcor^2 = 0$ denoting statistical independence.

Clinical symptom prediction in individuals with ASD

We investigated the relationship between stDNN-identified neurobiological features with the severity of clinical symptoms in individuals with ASD. The severity of clinical symptoms was

assessed using Autism Diagnostic Interview–Revised (ADI-R) domain scores. For this analysis, we combined the datasets from the ABIDE, Stanford, and GENDAAR cohorts (N_ASD = 417). The neurobiological features for each brain region were derived by computing the temporal median of the feature attribute weights which are obtained from the integrated gradients approach. The brain features were derived from each stDNN model obtained during the five-fold cross-validation. To find the relationship between the clinical symptoms with these brain features, we computed Spearman correlations between the ADI-R social scores and the brain features derived from each of the stDNN models. The aforementioned brain behavior analysis was repeated with ADI-R communication and ADI-R RRB scores. We applied FDR correction to correct for multiple ROI-level as well as ADI-R domain-level comparisons.

Supplementary Results

Results of control analyses: Classification of ASD vs Controls in the ABIDE cohort using conventional approaches

stDNN outperformed conventional approaches that used (i) static functional connectivity as features, which achieved an average accuracy of 63.1 ± 5.87%, an average precision of 0.64 ± 0.06, recall of 0.63 ± 0.07 and F1 score of 0.62 ± 0.06 (**Supplementary Table S5**), (ii) amplitude of low-frequency fluctuation (ALFF) as features, which achieved an average accuracy of 56.34 ± 1.19 %, an average precision of 0.57 ± 0.01, recall of 0.55 ± 0.04 and F1 score of 0.56 ± 0.02 (**Supplementary Table S6**), (iii) BOLD signal variability as features, which achieved an average accuracy of 58.77 ± 2.57%, an average precision of 0.60 ± 0.03, recall of 0.56 ± 0.05 and F1 score of 0.58 ± 0.04 (**Supplementary Table S7**), (iv) sliding-window functional connectivity variability as features, which achieved an average accuracy of 54.93 ± 2.55%, an average precision of 0.55 ± 0.03, recall of 0.58 ± 0.04 and F1 score of 0.56 ± 0.03 (**Supplementary Table S8**), (v) fMRI timeseries as features, which achieved an average accuracy of 50.27 ± 3.73 %, an average precision of 0.50 ± 0.04, recall of 0.50 ± 0.06 and F1 score of 0.50 ± 0.05 (**Supplementary Table S9**), further highlighting the strength of our approach

Results of control analyses: Classification of ASD vs Controls in the Stanford cohort using conventional approaches

stDNN outperformed conventional approaches that used (i) static functional connectivity as features, which achieved an average accuracy of 62.9 ± 5.58%, an average precision of 0.67 ± 0.09, recall of 0.55 ± 0.08 and F1 score of 0.60 ± 0.05 (**Supplementary Table S11**), (ii) ALFF as features, which achieved an average accuracy of 57.28 ± 2.26%, an average precision of 0.59 ± 0.03, recall of 0.54 ± 0.08 and F1 score of 0.55 ± 0.04 (**Supplementary Table S12**), (iii) BOLD signal variability as features, which achieved an average accuracy of 58.44 ± 2.89%, an average precision of 0.60 ± 0.04, recall of 0.57 ± 0.06 and F1 score of 0.57 ± 0.03 (**Supplementary Table S13**), (iv) sliding-window functional connectivity variability as features, which achieved an average accuracy of 59.07 ± 5.23%, an average precision of 0.59 ± 0.05, recall of 0.64 ± 0.06 and F1 score of 0.61 ± 0.04 (**Supplementary Table S14**), (v) fMRI timeseries as features, which achieved an average accuracy of 55.39 ± 1.43%, an average precision of 0.55 ± 0.01, recall of 0.58 ± 0.04 and F1 score of 0.56 ± 0.02 (**Supplementary Table S15**), further highlighting the strength of our approach.

Results of control analyses: Classification of ASD vs Controls in the GENDAAR cohort using conventional approaches

stDNN outperformed conventional approaches that used (i) static functional connectivity as features, which achieved an average accuracy of 52.57 ± 2.84%, an average precision of 0.51 ± 0.18, recall of 0.56 ± 0.20 and F1 score of 0.52 ± 0.18 (**Supplementary Table S17**), (ii) ALFF as features, which achieved an average accuracy of 54.57 ± 1.98%, an average precision of 0.49 ± 0.09, recall of 0.88 ± 0.16 and F1 score of 0.63 ± 0.13 (**Supplementary Table S18**), (iii) BOLD signal variability as features, which achieved an average accuracy of 54.79 ± 3.25%, an

average precision of 0.64 ± 0.05, recall of 0.45 ± 0.03 and F1 score of 0.52 ± 0.03 (**Supplementary Table S19**), (iv) sliding-window functional connectivity variability as features, which achieved an average accuracy of 59.71 ± 5.89%, an average precision of 0.63 ± 0.05, recall of 0.64 ± 0.06 and F1 score of 0.64 ± 0.06 (**Supplementary Table S20**), (v) fMRI timeseries as features, which achieved an average accuracy of 48.63 ± 3.14%, an average precision of 0.55 ± 0.03, recall of 0.47 ± 0.05 and F1 score of 0.50 ± 0.06 (**Supplementary Table S21**), further highlighting the strength of our approach.

Results of control analyses to demonstrate the robustness of the main findings to head motion-related confounds

First, we subsampled the two groups so that they were matched on head motion. 300 subjects in each group were randomly selected such that the two groups were matched on the mean framewise displacement. This procedure was repeated 100 times, yielding 100 datasets with 300 ASD subjects and 300 TD controls who were matched on head motion. We applied our five-fold cross-validation procedure on these datasets. Results of these analyses were consistent with the results from our original analysis: stDNN could distinguish between ASD and TD with a high accuracy of 76.2 ± 1.4% and a precision of 0.77 ± 0.02, recall of 0.76 ± 0.03 and F1 score of 0.76 ± 0.02, despite a significant reduction in the number of subjects (original N: 834, subsampled group N: 600).

Second, we examined head motion in correctly and incorrectly classified participants and found that the correctly classified participants and incorrectly classified participants did not differ significantly in mean framewise displacement (FD) ($p = 0.16$).

Third, we examined the univariate relationships between head motion as measured by mean FD and feature strength of stDNN-identified brain areas that contribute most significantly to ASD classification separately. We found no significant association between mean FD and feature strength in each of the stDNN-identified brain areas across the five folds ($p > 0.05$).

Fourth, we examined the multivariate relationships between head motion as measured by mean FD and feature strength of stDNN-identified brain areas that contribute most significantly to ASD classification. Specifically, we computed the squared distance correlation ($dcor^2$) between the strength of features and the mean framewise displacement (FD). Briefly, $dcor^2$ is a measure of the nonlinear relationship between multidimensional variables, making it a better measure than conventional metrics like Pearson correlation, which only capture univariate linear relationships. $dcor^2$ has a range from 0 to 1, with $dcor^2 = 0$ denoting statistical independence. We performed statistical testing to evaluate the hypothesis $dcor^2 = 0$. Specifically, we used the statistical procedures proposed by Szekely and colleagues (20) for assessing the significance of the correlation of distances, which were implemented in the Python package 'dcor'. We found no significant relationship between head motion and stDNN-identified brain features in each of the five folds ($dcor^2 = 0.03 \pm 0.004$, all ps > 0.1).

Fifth, we examined the relationship between stDNN-derived brain features and instantaneous head motion as measured by the percentage of scrubbed frames. We computed the squared distance correlation $dcor^2$ between the strength of features and the percentage of scrubbed frames (FD > 0.2 mm) and found no significant effect of head motion ($dcor^2 = 0.03 \pm 0.002$, all ps >

0.18).

Sixth, we appended a scrubbing stage to our original preprocessing pipeline. The volumes to be excised were first identified using the stringent criteria (FD > 0.2mm) recommended in the original Power et al. report. The excised volumes were reestimated using interpolation. We applied our five-fold cross-validation procedure to this 'scrubbed/interpolated' dataset. Results of these analyses were very similar to the results from our original analysis: stDNN could distinguish between ASD and TD with a high accuracy of 78 ± 3% and a precision of 0.76 ± 0.03, recall of 0.82 ± 0.03 and F1 score of 0.79 ± 0.03.

These results further attest to the robustness of our findings.

Results of control analysis to demonstrate that the accuracy of stDNN in distinguishing between ASD and TD is consistent across sites and sexes.

We determined site-wise classification accuracy by training stDNN model on data from all sites using a five-fold cross-validation setup, wherein we divided the entire dataset into five different parts: four parts were used for training and validation, with the fifth serving as the test set. Then, for each iteration, we rotate through the entire dataset five times, selecting a different section as the test set. We evaluate the performance of our stDNN model individually per site for each of the five subsets and then average the results across the five subsets to determine the mean and standard deviation values of the accuracy per site. We found no differences in classification accuracy across sites as evident from tight standard deviation bounds on classification accuracy

values across sites (78.18 ± 2.11%). We repeated the aforementioned analysis to determine sex-wise classification accuracy and found no differences in classification accuracy across sexes (male: 76.95%; female: 77.22%).

Results of error analysis: Comparing symptom severity of ASD participants correctly classified by stDNN to the symptom severity of ASD participants incorrectly classified

We compared the ADI-R scores of ASD participants who were correctly classified by our model to the ADI-R scores of ASD participants who were incorrectly classified. This analysis revealed that the ADI-R social ($p = 0.92$), ADI-R communication ($p = 0.53$), and the ADI-R restricted repetitive ($p = 0.19$) domain scores did not differ between the ASD participants correctly classified by our model and the ASD participants incorrectly classified.

Supplementary Discussion

stDNN advantages over commonly used DNN-based methods for classification of brain imaging data

stDNN overcomes the key limitations of commonly used DNN-based methods for classification of brain imaging data. First, unlike static connectivity-based approaches, stDNN does not assume the stationarity of fMRI timeseries and models the underlying non-stationarities in the data. Second, convolutional layers are parsimonious compared to fully connected layers in connectivity-based models. Third, temporal averaging, compared to the flattening layer in connectivity models, allows the examination of variable length fMRI time series data and with a comparatively smaller number of parameters to be trained. Fourth, stDNN is a "fully" convolutional approach, in which we can train on data of arbitrary length and test on, or generalize to, datasets of any arbitrary length. Additionally, our integrated XAI approach allows for simultaneous classification and identification of neurobiologically meaningful features that distinguish individuals with ASD from neurotypical controls.

Importantly, our approach also overcomes the weaknesses of neural network architectures such as recurrent neural networks, which have only been used for the analysis of single-site data ; their efficacy on multisite data as used in the present study is not known. Furthermore, although recurrent neural networks such as long short-term memory (LSTM) models can address the problem of vanishing and exploding gradients problem, they are difficult to train and are particularly problematic for training long time series from small groups of participants as is typically the case with fMRI data from neurodevelopmental populations (21). Critically, a study

using LSTMs to model fMRI timeseries showed poor decoding test accuracy, despite using one of the largest fMRI datasets from the Human Connectome project (22). Moreover, fMRI timeseries are dominated by short-range, rather than long-range, correlations (23-25). stDNN solves this problem and overcomes the limitations of extant methods by (i) simultaneously modelling latent spatiotemporal dynamics and dynamic interactions among distributed brain areas without any predefined assumptions or feature engineering, (ii) exploiting short-range spatiotemporal correlations using convolutional networks, (iii) discovering time invariant features, and (iv) using hierarchical features across network layers. Furthermore, unlike deep DNN models that have achieved success in computer vision applications, where a large number of labelled data are available, our stDNN model is relatively shallower. This is because the number of labelled datasets in fMRI applications such as ASD classification is limited, therefore fitting deep models, which have a significantly larger number of parameters to train than shallow models, can result in overfitting and poor test accuracies.

Limitations and future work

A few study limitations need to be taken into consideration and addressed by future studies. First, the current work leveraged large-scale publicly available datasets and examined latent spatiotemporal dynamics in brain-wide fMRI timeseries that were acquired using the most commonly used repetition time (TR = 2 sec). Future work is needed to investigate whether the study findings are present in fMRI data acquired using faster temporal sampling. Critically, our stDNN model trained on ABIDE cohort (TR = 2 sec) could accurately distinguish between ASD and TD controls without additional training in an independent GENDAAR cohort acquired at TR

= 1.6 sec, suggesting that our model and findings extend to relatively faster temporal sampling. Second, as in extant neuroimaging studies of ASD, our sample was predominantly male. Future research examining a much larger number of females is required to further establish the generalizability of our ASD classification and prediction model. Third, due to the relative advantages of the IG method over other XAI methods, the current work used IGs to identify brain features that distinguish between ASD and neurotypical controls. Further studies utilizing other XAI methods, which is an active and rapidly growing area of research, are needed to investigate whether the study findings are robust to feature attribution procedures.

Supplementary Tables

Supplementary Table S1. Summary of previous studies that have used deep learning and functional brain imaging data to distinguish ASD from controls.

Studies	Method	ABIDE Train/Test Protocol	ABIDE Result	Independent Cohort Testing	Symptom Prediction	Interpretable features
Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction(26)	Ensemble 3D CNN	10-fold cross validation - ABIDE I & Train ABD I and test ABD II	73.50%	No	No	No
Simple 1-D Convolutional Networks for Resting-State fMRI Based Classification in Autism(27)	1D CNN	Leave site out (ABD I+II)	64.00 + 0.1%	No	No	No
Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network(28)	sets of parallel 2D Conv blocks (2D CNN)	Leave one site out	70.20%	No	No	No
Identification of autism spectrum disorder using deep learning and the ABIDE dataset(29)	Denoising autoencoder	10-fold cross-val (stratified site wise)	70%	No	Yes	No
Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks(30)	LSTM models	10-fold cross-val (stratified site wise) 90/10 split	70.10%	No	No	No

Invertible Network for Classification and Biomarker Selection for ASD(31)	Invertible block (1D Net)	10-fold cross-val	71.10%	No	TBD	Yes
A Hybrid 3DCNN and 3DC-LSTM based model for 4D Spatio-temporal fMRI data: An ABIDE Autism Classification study(32)	Ensemble Mixture of 3D CNN and 3D Conv-LSTM models	5-fold cross-val	58.00 + 0.03%	No	No	No
Jointly Discriminative and Generative Recurrent Neural Networks for Learning from fMRI(33)	LSTM - DG Model	10-fold cross-val	Site wise results: No overall results	No	Yes	No
Detecting Abnormalities in Resting-State Dynamics: An Unsupervised Learning Approach(34)	Conv-LSTM encoder-decoder model	one time run (75/25 split)	Only NYU Site results	No	Yes	No
Functional connectivity magnetic resonance imaging classification of autism spectrum disorder using the multisite ABIDE dataset (35)	Combination of classical ML models	5-fold cross-val	71.98%	No	Yes	No
Autism Classification Using Topological Features and Deep Learning: A Cautionary Tale(36)	3-layer NN (concatenation)	5-fold cross-val (stratified)	69.20%	No	No	No

Supplementary Table S2. Demographic information for the ASD and TD Control groups in the ABIDE, Stanford and GENDAAR cohorts. In the ABIDE cohort, age, IQ and ADI data were available for 834, 821, and 295 participants, respectively. In the Stanford cohort, age, IQ and ADI data were available for 201, 198, and 75 participants, respectively. In the GENDAAR cohort, age, IQ and ADI data were available for 90, 0, and 47 participants, respectively.

	ABIDE		**Stanford**		**GENDAAR**	
	ASD (N=419)	TD (N=415)	ASD (N=101)	TD (N=101)	ASD (N=50)	TD (N=40)
Age (in years)	13.9 ± 6.6 (range: 5.2 to 42)	14.9 ± 5.9 (range: 5.9 to 40)	11.6 ± 2.3 (range: 7.5 to 18)	10.9 ± 2.4 (range: 7.4 to 16.6)	13.8 ± 2.9 (range: 8.2 to 18)	14.1 ± 2.6 (range: 8.9 to 18)
Gender (female\male)	51\368	75\340	16\85	18\83	10\40	10\30
IQ	105 ± 17	112 ± 14	112 ± 17	116 ± 14	NA	NA
ADI-R Social	18.8 ± 5.5		8.9 ± 4.0		19.7 ± 5.2	
ADI-R Verbal	15.4 ± 4 .5		8.8 ± 4.2		16.4 ± 4.2	
ADI-R Restricted and Repetitive Behavior	5.7 ± 2.6		4.0 ± 2.4		6.1 ± 2.8	

NA = data not available

Supplementary Table S3. ASD vs. TD classification and five-fold cross-validation accuracy in ABIDE cohort.

Fold Number	Accuracy	Precision	Recall	F1-score
Fold 1	75.45	0.74	0.80	0.77
Fold 2	74.85	0.72	0.81	0.76
Fold 3	82.63	0.80	0.88	0.84
Fold 4	78.44	0.77	0.82	0.79
Fold 5	79.52	0.79	0.81	0.80
Avg	**78.2 ± 2.84 %**	**0.76 ± 0.03**	**0.82 ± 0.03**	**0.79 ± 0.03**

Supplementary Table S4. ASD vs. TD classification and five-fold cross-validation accuracy in ABIDE cohort using other commonly-used brain atlases.

Atlas	No of ROIs	Accuracy	Precision	Recall	F1-score
Schaefer (37)	1000	78.78 ± 2.47 %	0.80 ± 0.02	0.77 ± 0.05	0.78 ± 0.03
CC200 (38)	200	78.42 ± 2.37 %	0.79 ± 0.03	0.78 ± 0.02	0.78 ± 0.01
CC400 (38)	392	78.18 ± 2.08 %	0.77 ± 0.04	0.81 ± 0.04	0.79 ± 0.02
Brainnetome atlas	**246**	**78.18 ± 2.84 %**	**0.76 ± 0.03**	**0.82 ± 0.03**	**0.79 ± 0.03**

Supplementary Table S5. ASD vs. TD classification and five-fold cross-validation accuracies in ABIDE cohort using static functional connectivity as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	58.51 ± 1.13 %	0.60 ± 0.01	0.51 ± 0.03	0.55 ± 0.02
Decision Tree	53.00 ± 2.48 %	0.53 ± 0.02	0.57 ± 0.06	0.55 ± 0.04
Linear SVM	67.75 ± 3.21 %	0.68 ± 0.04	0.68 ± 0.03	0.68 ± 0.03
Logistic Reg	69.91 ± 3.66 %	0.70 ± 0.04	0.70 ± 0.03	0.70 ± 0.03
Ridge Classifier	66.67 ± 3.78 %	0.67 ± 0.04	0.67 ± 0.04	0.67 ± 0.04
LASSO	64.51 ± 4.28 %	0.65 ± 0.04	0.64 ± 0.08	0.64 ± 0.06
Random Forest	62.12 ± 5.07 %	0.62 ± 0.05	0.62 ± 0.07	0.62 ± 0.05
Avg	**63.12 ± 5.87 %**	**0.64 ± 0.06**	**0.63 ± 0.07**	**0.62 ± 0.06**
Our stDNN model Avg (see Supplementary Table S3)	**78.18 ± 2.84 %**	**0.76 ± 0.03**	**0.82 ± 0.03**	**0.79 ± 0.03**

Supplementary Table S6. ASD vs. TD classification and five-fold cross-validation accuracies in ABIDE cohort using amplitude of low-frequency fluctuation (ALFF) as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	55.51 ± 1.06 %	0.57 ± 0.01	0.48 ± 0.03	0.52 ± 0.02
Decision Tree	54.80 ± 3.68 %	0.55 ± 0.03	0.59 ± 0.06	0.57 ± 0.04
Linear SVM	56.96 ± 4.55 %	0.58 ± 0.05	0.55 ± 0.05	0.56 ± 0.04
Logistic Reg	57.80 ± 2.35 %	0.58 ± 0.03	0.57 ± 0.04	0.57 ± 0.02
Ridge Classifier	56.61 ± 2.87 %	0.57 ± 0.04	0.56 ± 0.03	0.57 ± 0.02
LASSO	56.96 ± 3.68 %	0.58 ± 0.04	0.57 ± 0.04	0.57 ± 0.03
Random Forest	57.07 ± 1.85 %	0.57 ± 0.02	0.58 ± 0.05	0.58 ± 0.03
Avg	**56.34 ± 1.19 %**	**0.57 ± 0.01**	**0.55 ± 0.04**	**0.56 ± 0.02**
Our stDNN model Avg (see Supplementary Table S3)	**78.18 ± 2.84 %**	**0.76 ± 0.03**	**0.82 ± 0.03**	**0.79 ± 0.03**

Supplementary Table S7. ASD vs. TD classification and five-fold cross-validation accuracies in ABIDE cohort using BOLD signal variability as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	55.88 ± 3.36 %	0.57 ± 0.04	0.50 ±0.03	0.53 ± 0.03
Decision Tree	55.40 ± 2.97 %	0.56 ±0.03	0.55 ±0.04	0.55 ± 0.03
Linear SVM	57.91 ± 2.96 %	0.59 ±0.04	0.52 ±0.07	0.55 ± 0.05
Logistic Reg	62.22 ± 3.87 %	0.63 ±0.04	0.61 ±0.07	0.62 ± 0.05
Ridge Classifier	58.63 ± 2.79 %	0.60 ±0.03	0.55 ±0.07	0.57 ± 0.05
LASSO	60.67 ± 2.64 %	0.62 ±0.02	0.56 ±0.08	0.59 ± 0.05
Random Forest	60.67 ± 2.88 %	0.60 ±0.02	0.64 ±0.06	0.62 ± 0.04
Avg	**58.77 ± 2.57 %**	**0.60 ± 0.03**	**0.56 ± 0.05**	**0.58 ± 0.04**
Our stDNN model Avg (see Supplementary Table S3)	**78.18 ± 2.84 %**	**0.76 ± 0.03**	**0.82 ± 0.03**	**0.79 ± 0.03**

Supplementary Table S8. ASD vs. TD classification and five-fold cross-validation accuracies in ABIDE cohort using sliding-window functional connectivity variability as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	53.71 ± 1.75 %	0.53 ± 0.01	0.66 ± 0.08	0.59 ± 0.03
Decision Tree	49.76 ± 2.37 %	0.50 ± 0.02	0.52 ± 0.02	0.51 ± 0.02
Linear SVM	56.35 ± 1.44 %	0.57 ± 0.01	0.56 ± 0.07	0.56 ± 0.03
Logistic Reg	56.83 ± 1.01 %	0.57 ± 0.01	0.59 ± 0.05	0.58 ± 0.02
Ridge Classifier	57.07 ± 1.25 %	0.57 ± 0.01	0.58 ± 0.07	0.57 ± 0.03
LASSO	55.75 ± 2.48 %	0.56 ± 0.03	0.55 ± 0.05	0.55 ± 0.03
Random Forest	55.03 ± 3.44 %	0.55 ± 0.03	0.58 ± 0.02	0.56 ± 0.02
Avg	**54.93 ± 2.55 %**	**0.55 ± 0.03**	**0.58 ± 0.04**	**0.56 ± 0.03**
Our stDNN model Avg (see Supplementary Table S3)	**78.18 ± 2.84 %**	**0.76 ± 0.03**	**0.82 ± 0.03**	**0.79 ± 0.03**

Supplementary Table S9. ASD vs. TD classification and five-fold cross-validation accuracies in ABIDE cohort using fMRI timeseries as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	51.44 ± 3.64 %	0.51 ± 0.03	0.54 ± 0.06	0.53 ± 0.05
Decision Tree	54.31 ± 3.93 %	0.55 ± 0.04	0.53 ± 0.05	0.54 ± 0.04
Linear SVM	48.68 ± 2.69 %	0.49 ± 0.03	0.48 ± 0.05	0.48 ± 0.04
Logistic Reg	48.55 ± 4.08 %	0.49 ± 0.04	0.48 ± 0.04	0.48 ± 0.04
Ridge Classifier	48.56 ± 4.61 %	0.49 ± 0.05	0.48 ± 0.08	0.48 ± 0.06
LASSO	50.84 ± 2.34 %	0.51 ± 0.03	0.49 ± 0.06	0.50 ± 0.04
Random Forest	49.52 ± 4.86 %	0.49 ± 0.04	0.53 ± 0.09	0.51 ± 0.06
Avg	**50.27 ± 3.73 %**	**0.50 ± 0.04**	**0.50 ± 0.06**	**0.50 ± 0.05**
Our stDNN model Avg (see Supplementary Table S3)	**78.18 ± 2.84 %**	**0.76 ± 0.03**	**0.82 ± 0.03**	**0.79 ± 0.03**

Supplementary Table S10. ASD vs. TD classification and five-fold cross-validation accuracy in Stanford cohort.

Fold Number	Accuracy	Precision	Recall	F1-score
Fold 1	73.76	0.71	0.79	0.75
Fold 2	71.78	0.72	0.72	0.72
Fold 3	68.81	0.64	0.88	0.74
Fold 4	74.75	0.69	0.89	0.78
Fold 5	73.76	0.72	0.77	0.75
Avg	**72.57 ± 2.11%**	**0.70 ± 0.03**	**0.81 ± 0.06**	**0.75 ± 0.02**

Supplementary Table S11. ASD vs. TD classification and five-fold cross-validation accuracies in Stanford cohort using static functional connectivity as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	57.82 ± 1.10 %	0.60 ± 0.02	0.50 ± 0.03	0.54 ± 0.01
Decision Tree	52.77 ± 2.78 %	0.53 ± 0.03	0.53 ± 0.07	0.53 ± 0.04
Linear SVM	66.93 ± 1.91 %	0.74 ± 0.04	0.53 ± 0.05	0.62 ± 0.03
Logistic Reg	67.52 ± 2.23 %	0.74 ± 0.04	0.54 ± 0.05	0.62 ± 0.04
Ridge Classifier	65.54 ± 2.07 %	0.72 ± 0.04	0.51 ± 0.05	0.60 ± 0.04
LASSO	66.63 ± 1.58 %	0.74 ± 0.04	0.52 ± 0.05	0.61 ± 0.03
Random Forest	62.77 ± 3.33 %	0.61 ± 0.03	0.74 ± 0.03	0.67 ± 0.02
Avg	**62.85 ± 5.58 %**	**0.67 ± 0.09**	**0.55 ± 0.08**	**0.60 ± 0.05**
Our stDNN model Avg (see Supplementary Table S10)	**72.57 ± 2.11%**	**0.70 ± 0.03**	**0.81 ± 0.06**	**0.75 ± 0.02**

Supplementary Table S12. ASD vs. TD classification and five-fold cross-validation accuracies in Stanford cohort using amplitude of low-frequency fluctuation (ALFF) as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	57.23 ± 3.45 %	0.59 ± 0.04	0.53 ± 0.07	0.55 ± 0.02
Decision Tree	52.97 ± 2.56 %	0.54 ± 0.03	0.52 ± 0.09	0.52 ± 0.04
Linear SVM	55.94 ± 3.86 %	0.57 ± 0.04	0.48 ± 0.17	0.51 ± 0.10
Logistic Reg	58.71 ± 1.89 %	0.63 ± 0.04	0.45 ± 0.10	0.52 ± 0.06
Ridge Classifier	59.70 ± 1.67 %	0.61 ± 0.02	0.56 ± 0.09	0.58 ± 0.05
LASSO	58.81 ± 3.35 %	0.60 ± 0.03	0.55 ± 0.15	0.56 ± 0.08
Random Forest	57.62 ± 2.23 %	0.56 ± 0.02	0.69 ± 0.04	0.62 ± 0.01
Avg	**57.28 ± 2.26 %**	**0.59 ± 0.03**	**0.54 ± 0.08**	**0.55 ± 0.04**
Our stDNN model Avg (see Supplementary Table S10)	**72.57 ± 2.11%**	**0.70 ± 0.03**	**0.81 ± 0.06**	**0.75 ± 0.02**

Supplementary Table S13. ASD vs. TD classification and five-fold cross-validation accuracies in Stanford cohort using BOLD signal variability as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	59.01 ± 1.75 %	0.62 ± 0.03	0.49 ± 0.08	0.54 ± 0.04
Decision Tree	52.48 ± 3.58 %	0.53 ± 0.04	0.52 ± 0.09	0.52 ± 0.04
Linear SVM	58.02 ± 1.75 %	0.58 ± 0.02	0.62 ± 0.04	0.59 ± 0.02
Logistic Reg	61.68 ± 2.21 %	0.64 ± 0.03	0.55 ± 0.06	0.59 ± 0.03
Ridge Classifier	58.71 ± 3.45 %	0.60 ± 0.05	0.59 ± 0.14	0.58 ± 0.07
LASSO	60.30 ± 0.96 %	0.62 ± 0.02	0.55 ± 0.09	0.58 ± 0.04
Random Forest	58.91 ± 2.26 %	0.59 ± 0.03	0.65 ± 0.08	0.61 ± 0.01
Avg	**58.44 ± 2.89 %**	**0.60 ± 0.04**	**0.57 ± 0.06**	**0.57 ± 0.03**
Our stDNN model Avg (see Supplementary Table S10)	**72.57 ± 2.11%**	**0.70 ± 0.03**	**0.81 ± 0.06**	**0.75 ± 0.02**

Supplementary Table S14. ASD vs. TD classification and five-fold cross-validation accuracies in Stanford cohort using sliding-window functional connectivity variability as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	53.27 ± 2.36 %	0.53 ± 0.02	0.70 ± 0.04	0.60 ± 0.02
Decision Tree	52.77 ± 1.24 %	0.53 ± 0.01	0.55 ± 0.06	0.54 ± 0.03
Linear SVM	63.66 ± 1.35 %	0.63 ± 0.02	0.66 ± 0.04	0.65 ± 0.02
Logistic Reg	64.85 ± 2.54 %	0.64 ± 0.02	0.70 ± 0.05	0.66 ± 0.03
Ridge Classifier	63.27 ± 1.01 %	0.63 ± 0.01	0.66 ± 0.04	0.64 ± 0.01
LASSO	60.59 ± 1.31 %	0.61 ± 0.01	0.61 ± 0.05	0.60 ± 0.03
Random Forest	55.05 ± 3.79 %	0.55 ± 0.04	0.59 ± 0.06	0.57 ± 0.03
Avg	**59.07 ± 5.23 %**	**0.59 ± 0.05**	**0.64 ± 0.06**	**0.61 ± 0.04**
Our stDNN model Avg (see Supplementary Table S10)	**72.57 ± 2.11 %**	**0.70 ± 0.03**	**0.81 ± 0.06**	**0.75 ± 0.02**

Supplementary Table S15. ASD vs. TD classification and five-fold cross-validation accuracies in Stanford cohort using fMRI timeseries as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	53.66 ± 0.97 %	0.53 ± 0.01	0.69 ± 0.02	0.60 ± 0.01
Decision Tree	53.76 ± 2.36 %	0.54 ± 0.02	0.52 ± 0.05	0.53 ± 0.03
Linear SVM	57.03 ± 0.73 %	0.57 ± 0.01	0.56 ± 0.03	0.56 ± 0.02
Logistic Reg	59.60 ± 0.97 %	0.60 ± 0.01	0.60 ± 0.03	0.60 ± 0.01
Ridge Classifier	55.54 ± 1.45 %	0.56 ± 0.01	0.54 ± 0.03	0.55 ± 0.02
LASSO	55.05 ± 1.64%	0.55 ± 0.02	0.53 ± 0.04	0.54 ± 0.03
Random Forest	53.07 ± 1.91 %	0.53 ± 0.02	0.62 ± 0.06	0.57 ± 0.02
Avg	**55.39 ± 1.43 %**	**0.55 ± 0.01**	**0.58 ± 0.04**	**0.56 ± 0.02**
Our stDNN model Avg (see Supplementary Table S10)	**72.57 ± 2.11 %**	**0.70 ± 0.03**	**0.81 ± 0.06**	**0.75 ± 0.02**

Supplementary Table S16. ASD vs. TD classification and five-fold cross-validation accuracy in GENDAAR cohort.

Fold Number	Accuracy	Precision	Recall	F1-score
Fold 1	78.89	0.78	0.86	0.82
Fold 2	77.78	0.80	0.80	0.80
Fold 3	71.11	0.72	0.78	0.75
Fold 4	73.33	0.74	0.80	0.77
Fold 5	74.44	0.77	0.78	0.77
Avg	**75.11 ± 2.86%**	**0.76 ± 0.03**	**0.80 ± 0.03**	**0.78 ± 0.02**

Supplementary Table S17. ASD vs. TD classification and five-fold cross-validation accuracies in GENDAAR cohort using static functional connectivity as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	52.67 ± 3.95 %	0.59 ± 0.05	0.47 ± 0.04	0.52 ± 0.04
Decision Tree	52.22 ± 2.72 %	0.57 ± 0.02	0.54 ± 0.11	0.55 ± 0.06
Linear SVM	46.67 ± 4.44 %	0.11 ± 0.22	0.20 ± 0.40	0.14 ± 0.29
Logistic Reg	55.11 ± 0.89 %	0.57 ± 0.01	0.80 ± 0.01	0.66 ± 0.01
Ridge Classifier	54.67 ± 1.78 %	0.57 ± 0.01	0.78 ± 0.05	0.65 ± 0.02
LASSO	52.44 ± 1.91 %	0.60 ± 0.04	0.50 ± 0.21	0.51 ± 0.12
Random Forest	54.22 ± 5.32 %	0.58 ± 0.04	0.61 ± 0.16	0.58 ± 0.10
Avg	**52.57 ± 2.84 %**	**0.51 ± 0.18**	**0.56 ± 0.20**	**0.52 ± 0.18**
Our stDNN model Avg (see Supplementary Table S16)	**75.11 ± 2.86 %**	**0.76 ± 0.03**	**0.80 ± 0.03**	**0.78 ± 0.02**

Supplementary Table S18. ASD vs. TD classification and five-fold cross-validation accuracies in GENDAAR cohort using amplitude of low-frequency fluctuation (ALFF) as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	51.11 ± 5.44 %	0.33 ± 0.27	0.60 ± 0.49	0.43 ± 0.35
Decision Tree	53.56 ± 4.58 %	0.45 ± 0.22	0.77 ± 0.39	0.57 ± 0.28
Linear SVM	56.67 ± 0.99 %	0.56 ± 0.01	1.00 ± 0.01	0.72 ± 0.00
Logistic Reg	55.56 ± 0.00 %	0.56 ± 0.00	1.00 ± 0.00	0.71 ± 0.00
Ridge Classifier	55.56 ± 0.70 %	0.56 ± 0.00	1.00 ± 0.01	0.71 ± 0.00
LASSO	56.22 ± 1.33 %	0.56 ± 0.01	0.98 ± 0.03	0.71 ± 0.00
Random Forest	53.33 ± 4.44 %	0.44 ± 0.22	0.80 ± 0.40	0.57 ± 0.29
Avg	**54.57 ± 1.98 %**	**0.49 ± 0.09**	**0.88 ± 0.16**	**0.63 ± 0.13**
Our stDNN model Avg (see Supplementary Table S16)	**75.11 ± 2.86 %**	**0.76 ± 0.03**	**0.80 ± 0.03**	**0.78 ± 0.02**

Supplementary Table S19. ASD vs. TD classification and five-fold cross-validation accuracies in GENDAAR cohort using BOLD signal variability as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	50.67 ± 3.49 %	0.58 ± 0.05	0.42 ± 0.03	0.48 ± 0.03
Decision Tree	58.44 ± 6.80 %	0.67 ± 0.07	0.50 ± 0.08	0.57 ± 0.07
Linear SVM	56.00 ± 3.56 %	0.66 ± 0.07	0.44 ± 0.07	0.52 ± 0.05
Logistic Reg	56.00 ± 4.37 %	0.65 ± 0.06	0.46 ± 0.06	0.54 ± 0.05
Ridge Classifier	49.78 ± 2.57 %	0.56 ± 0.04	0.46 ± 0.11	0.50 ± 0.06
LASSO	56.67 ± 4.71 %	0.66 ± 0.05	0.46 ± 0.09	0.53 ± 0.07
Random Forest	56.00 ± 4.07 %	0.67 ± 0.04	0.41 ± 0.01	0.50 ± 0.08
Avg	**54.79 ± 3.25 %**	**0.64 ± 0.05**	**0.45 ± 0.03**	**0.52 ± 0.03**
Our stDNN model Avg (see Supplementary Table S16)	**75.11 ± 2.86 %**	**0.76 ± 0.03**	**0.80 ± 0.03**	**0.78 ± 0.02**

Supplementary Table S20. ASD vs. TD classification and five-fold cross-validation accuracies in GENDAAR cohort using sliding-window functional connectivity variability as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	56.22 ± 4.13 %	0.61 ± 0.04	0.58 ± 0.08	0.59 ± 0.05
Decision Tree	52.00 ± 4.63 %	0.57 ± 0.04	0.55 ± 0.07	0.56 ± 0.05
Linear SVM	65.78 ± 2.76 %	0.69 ± 0.02	0.70 ± 0.03	0.69 ± 0.03
Logistic Reg	65.33 ± 2.37 %	0.68 ± 0.02	0.70 ± 0.03	0.69 ± 0.02
Ridge Classifier	66.00 ± 3.19 %	0.69 ± 0.02	0.70 ± 0.04	0.70 ± 0.03
LASSO	58.00 ± 3.40 %	0.62 ± 0.03	0.63 ± 0.03	0.63 ± 0.03
Random Forest	54.67 ± 6.68 %	0.58 ± 0.05	0.62 ± 0.08	0.60 ± 0.07
Avg	**59.71 ± 5.89 %**	**0.63 ± 0.05**	**0.64 ± 0.06**	**0.64 ± 0.06**
Our stDNN model Avg (see Supplementary Table S16)	**75.11 ± 2.86 %**	**0.76 ± 0.03**	**0.80 ± 0.03**	**0.78 ± 0.02**

Supplementary Table S21. ASD vs. TD classification and five-fold cross-validation accuracies in GENDAAR cohort using fMRI timeseries as features.

Classifier	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	47.33 ± 3.41 %	0.52 ± 0.02	0.66 ± 0.05	0.58 ± 0.04
Decision Tree	48.00 ± 4.63 %	0.55 ± 0.06	0.42 ± 0.03	0.48 ± 0.02
Linear SVM	49.33 ± 1.94 %	0.55 ± 0.02	0.45 ± 0.02	0.50 ± 0.02
Logistic Reg	49.33 ± 4.80 %	0.56 ± 0.06	0.40 ± 0.04	0.47 ± 0.05
Ridge Classifier	48.44 ± 2.39 %	0.54 ± 0.02	0.46 ± 0.04	0.50 ± 0.03
LASSO	50.44 ± 2.06 %	0.57 ± 0.02	0.41 ±0.03	0.48 ± 0.03
Random Forest	47.56 ± 2.76 %	0.53 ± 0.03	0.46 ± 0.14	0.48 ± 0.09
Avg	**48.63 ± 3.14 %**	**0.55 ± 0.03**	**0.47 ± 0.05**	**0.50 ± 0.06**
Our stDNN model Avg (see Supplementary Table S16)	**75.11 ± 2.86 %**	**0.76 ± 0.03**	**0.80 ± 0.03**	**0.78 ± 0.02**

Supplementary Table S22. Demographic information for the ADHD and TD Control groups in the ADHD200-NYU cohort.

	ADHD200-NYU	
	ADHD (N=123)	TD (N=98)
Age (in years)	11.2 ± 2.7 (range: 7.2 to 17.6)	12.3 ± 3.0 (range: 7.2 to 18)
Gender (female\male)	34\89	48\50

Supplementary Table S23. ADHD vs. TD classification and five-fold cross-validation accuracy in ADHD200-NYU cohort.

Fold Number	Accuracy	Precision	Recall	F1-score
Fold 1	54.30	0.68	0.34	0.45
Fold 2	57.92	0.68	0.46	0.55
Fold 3	53.85	0.70	0.31	0.42
Fold 4	58.37	0.67	0.49	0.57
Fold 5	55.66	0.70	0.35	0.47
Avg	**56.0 ± 1.84%**	**0.69 ± 0.01**	**0.39 ± 0.07**	**0.49 ± 0.06**

Supplementary Table S24. Brain regions underlying ASD classification with the highest classification feature attributes (top 5%) in the ABIDE cohort.

Brain Region	Subdivision	Brainnetome Atlas ID & Region Label	Feature Attribution Weights (10^{-3})
Prefrontal Cortex	A9/46d, dorsal area 9/46	(16), MFG_R_7_1	2
Prefrontal Cortex	A9/46v, ventral area 9/46	(21), MFG_L_7_4	1.4
Prefrontal Cortex	A9/46v, ventral area 9/46	(22), MFG_R_7_4	1.7
Prefrontal Cortex	A10l, lateral area 10	(27), MFG_L_7_7	1.5
Prefrontal Cortex	IFS, inferior frontal sulcus	(32), IFG_R_6_2	1.4
Prefrontal Cortex	A45c, caudal area 45	(34), IFG_R_6_3	1.5
Prefrontal Cortex	A44v, ventral area 44	(39), IFG_L_6_6	1.4
Prefrontal Cortex	A14m, medial area 14	(41), OrG_L_6_1	1.5
Superior Temporal Gyrus	rpSTS, rostroposterior superior temporal sulcus	(122), pSTS_R_2_1	1.8
Precuneus	A7m, medial area 7 (PEp)	(147), Pcun_L_4_1	1.6
Posterior Cingulate Cortex	A23d, dorsal area 23	(175), CG_L_7_1	1.6
Posterior Cingulate Cortex	A23d, dorsal area 23	(176), CG_R_7_1	2.3
Posterior Cingulate Cortex	A23v, ventral area 23	(182), CG_R_7_4	1.4

Supplementary Table S25. Brain regions underlying ASD classification with the highest classification feature attributes (top 5%) in the Stanford cohort.

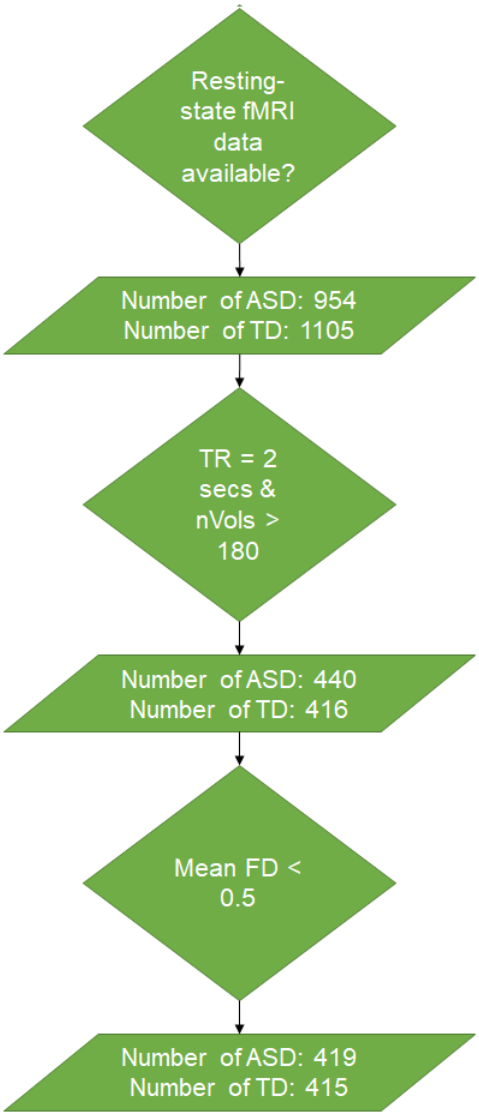
Brain Region	Subdivision	Brainnetome Atlas ID & Region Label	Feature Attribution Weights (10^{-3})
Prefrontal Cortex	A9/46d, dorsal area 9/46	(16), MFG_R_7_1	6.8
Prefrontal Cortex	A9/46v, ventral area 9/46	(21), MFG_L_7_4	4.3
Prefrontal Cortex	A9/46v, ventral area 9/46	(22), MFG_R_7_4	6
Prefrontal Cortex	A10l, lateral area 10	(27), MFG_L_7_7	5.9
Prefrontal Cortex	A44v, ventral area 44	(39), IFG_L_6_6	5.5
Superior Temporal Gyrus	rpSTS, rostroposterior superior temporal sulcus	(122), pSTS_R_2_1	4.7
Superior Temporal Gyrus	cpSTS, caudoposterior superior temporal sulcus	(124), pSTS_R_2_2	4.7
Superior Parietal Lobule	A7c, caudal area 7	(127), SPL_L_5_2	6.3
Precuneus	A7m, medial area 7 (PEp)	(147), Pcun_L_4_1	8.3
Precuneus	A5m, medial area 5 (PEm)	(149), Pcun_L_4_2	4.8
Posterior Cingulate Cortex	A23d, dorsal area 23	(175), CG_L_7_1	4.7
Posterior Cingulate Cortex	A23d, dorsal area 23	(176), CG_R_7_1	7.2
Cuneus	cLinG, caudal lingual gyrus	(190), Cun_R_5_1	5.4

Supplementary Table S26. Brain regions underlying ASD classification with the highest classification feature attributes (top 5%) in the GENDAAR cohort.

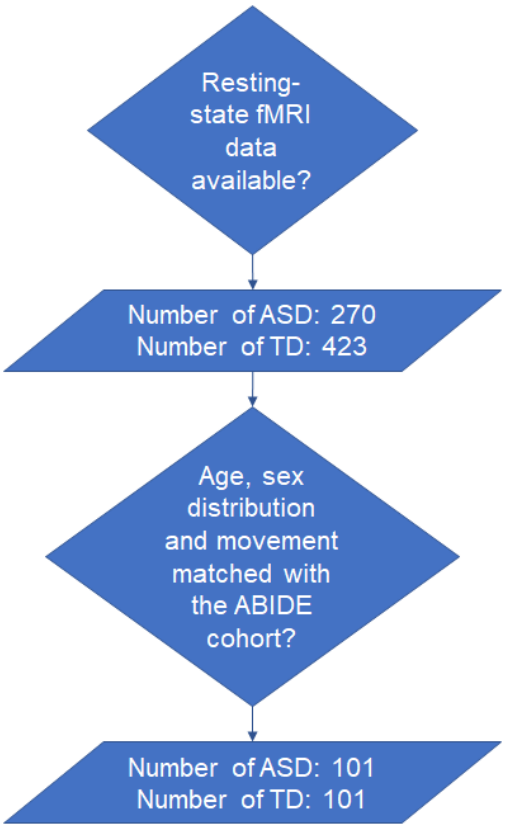
Brain Region	Subdivision	Brainnetome Atlas ID & Region Label	Feature Attribution Weights (10^{-3})
Prefrontal Cortex	A9/46d, dorsal area 9/46,Right	(16), MFG_R_7_1	0.94
Prefrontal Cortex	A9/46v, ventral area 9/46 ,Left	(21), MFG_L_7_4	0.90
Prefrontal Cortex	IFS, inferior frontal sulcus,Right	(32), IFG_R_6_2	1.02
Superior Temporal Gyrus	A38l, lateral area 38,Right	(78), STG_R_6_5	0.95
Fusiform Gyrus	A37mv, medioventral area37,Right	(106), FuG_R_3_2	0.90
Superior Temporal Gyrus	rpSTS, rostroposterior superior temporal sulcus,Right	(122), pSTS_R_2_1	1.18
Superior Parietal Lobule	A7c, caudal area 7,Left	(127), SPL_L_5_2	1.44
Superior Parietal Lobule	A5l, lateral area 5,Left	(129), SPL_L_5_3	1.09
Precuneus	A7m, medial area 7(PEp),Left	(147), PCun_L_4_1	1.04
Precuneus	A5m, medial area 5(PEm),Left	(149), Pcun_L_4_2	0.91
Posterior Cingulate Cortex	A23d, dorsal area 23,Right	(176), CG_R_7_1	1.17
Cuneus	cCunG, caudal cuneus gyrus,Right	(194), MVOcC _R_5_3	1.46

Supplementary Figures

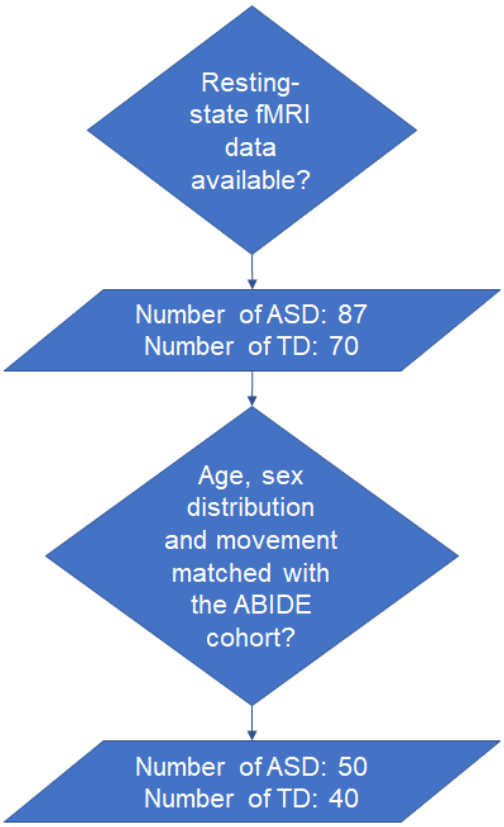
Supplementary Figure S1. Participant selection procedure for the ABIDE cohort.



Supplementary Figure S2. Participant selection procedure for the Stanford cohort.



Supplementary Figure S3. Participant selection procedure for the GENDAAR cohort.



Supplementary References

1. Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, et al. (2014): The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry*. 19:659-667.

2. Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson JS, Assaf M, et al. (2017): Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci Data*. 4:170010.

3. Abrams DA, Padmanabhan A, Chen T, Odriozola P, Baker AE, Kochalka J, et al. (2019): Impaired voice processing in reward and salience circuits predicts social communication in children with autism. *Elife*. 8.

4. Abrams DA, Lynch CJ, Cheng KM, Phillips J, Supekar K, Ryali S, et al. (2013): Underconnectivity between voice-selective cortex and reward circuitry in children with autism. *Proc Natl Acad Sci U S A*. 110:12060-12065.

5. Lynch CJ, Uddin LQ, Supekar K, Khouzam A, Phillips J, Menon V (2013): Default mode network in childhood autism: posteromedial cortex heterogeneity and relationship with social deficits. *Biol Psychiatry*. 74:212-219.

6. Uddin LQ, Supekar K, Lynch CJ, Khouzam A, Phillips J, Feinstein C, et al. (2013): Salience network-based classification and prediction of symptom severity in children with autism. *JAMA Psychiatry*. 70:869-879.

7. Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, et al. (2016): The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cereb Cortex*. 26:3508-3526.

8. Davatzikos C (2019): Machine learning in neuroimaging: Progress and challenges. *Neuroimage*. 197:652-656.

9. Kingma DP, Ba J (2015): Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. 1412.

10. van den Heuvel MP, Hulshoff Pol HE (2010): Exploring the brain network: a review on resting-state fMRI functional connectivity. *Eur Neuropsychopharmacol*. 20:519-534.

11. Zang YF, He Y, Zhu CZ, Cao QJ, Sui MQ, Liang M, et al. (2007): Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain Dev*. 29:83-91.

12. Garrett DD, Kovacevic N, McIntosh AR, Grady CL (2010): Blood oxygen level-dependent signal variability is more than just noise. *J Neurosci*. 30:4914-4921.

13. Allen EA, Damaraju E, Plis SM, Erhardt EB, Eichele T, Calhoun VD (2014): Tracking whole-brain connectivity dynamics in the resting state. *Cereb Cortex*. 24:663-676.

14. Lundberg SM, Lee S-I (2017): A unified approach to interpreting model predictions. *Advances in neural information processing systems*, pp 4765-4774.

15. Simonyan K, Vedaldi A, Zisserman A (2013): Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:13126034*.

16. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2014): Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:14126806*.

17. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017): Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, pp 618-626.

18. Sundararajan M, Taly A, Yan Q (2017): Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*: JMLR. org, pp 3319-3328.

19. Székely GJ, Rizzo ML, Bakirov NK (2007): Measuring and testing dependence by correlation of distances. *The annals of statistics*. 35:2769-2794.

20. Székely GJ, Rizzo ML, Bakirov NK (2007): Measuring and testing dependence by correlation of distances. *The annals of statistics*. 35:2769-2794.

21. Gers FA, Schmidhuber J, Cummins F (2000): Learning to forget: continual prediction with LSTM. *Neural Comput*. 12:2451-2471.

22. Thomas AW, Heekeren HR, Muller KR, Samek W (2019): Analyzing Neuroimaging Data Through Recurrent Deep Learning Models. *Front Neurosci*. 13:1321.

23. Ryali S, Supekar K, Chen T, Kochalka J, Cai W, Nicholas J, et al. (2016): Temporal Dynamics and Developmental Maturation of Salience, Default and Central-Executive Network Interactions Revealed by Variational Bayes Hidden Markov Modeling. *PLoS Comput Biol*. 12:e1005138.

24. Taghia J, Cai W, Ryali S, Kochalka J, Nicholas J, Chen T, et al. (2018): Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nat Commun*. 9:2505.

25. Ryali S, Supekar K, Chen T, Menon V (2011): Multivariate dynamical systems models for estimating causal interactions in fMRI. *Neuroimage*. 54:807-823.

26. Khosla M, Jamison K, Kuceyeski A, Sabuncu MR (2019): Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. *Neuroimage*. 199:651-662.

27. El Gazzar A, Cerliani L, van Wingen G, Thomas RM (2019): Simple 1-D convolutional networks for resting-state fMRI based classification in autism. *2019 International Joint Conference on Neural Networks (IJCNN)*: IEEE, pp 1-6.

28. Sherkatghanad Z, Akhondzadeh M, Salari S, Zomorodi-Moghadam M, Abdar M, Acharya UR, et al. (2019): Automated detection of autism spectrum disorder using a convolutional neural network. *Frontiers in Neuroscience*. 13.

29. Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F (2018): Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuroimage Clin*. 17:16-23.

30. Dvornek NC, Ventola P, Duncan JS (2018): Combining Phenotypic and Resting-State Fmri Data for Autism Classification with Recurrent Neural Networks. *Proc IEEE Int Symp Biomed Imaging*. 2018:725-728.

31. Zhuang J, Dvornek NC, Li X, Ventola P, Duncan JS (2019): Invertible network for classification and biomarker selection for asd. *International Conference on Medical Image Computing and Computer-Assisted Intervention*: Springer, pp 700-708.

32. El-Gazzar A, Quaak M, Cerliani L, Bloem P, van Wingen G, Thomas RM (2019): A Hybrid 3DCNN and 3DC-LSTM based model for 4D Spatio-temporal fMRI data: An ABIDE Autism Classification study. *OR 20 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*: Springer, pp 95-102.

33. Dvornek NC, Li X, Zhuang J, Duncan JS (2019): Jointly discriminative and generative recurrent neural networks for learning from fmri. *International Workshop on Machine Learning in Medical Imaging*: Springer, pp 382-390.

34. Khosla M, Jamison K, Kuceyeski A, Sabuncu MR (2019): Detecting Abnormalities in Resting-State Dynamics: An Unsupervised Learning Approach. Cham: Springer International Publishing, pp 301-309.

35. Yang X, Islam MS, Khaled AA (2019): Functional connectivity magnetic resonance imaging classification of autism spectrum disorder using the multisite ABIDE dataset. *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*: IEEE, pp 1-4.

36. Rathore A, Palande S, Anderson JS, Zielinski BA, Fletcher PT, Wang B (2019): Autism Classification Using Topological Features and Deep Learning: A Cautionary Tale. *International Conference on Medical Image Computing and Computer-Assisted Intervention*: Springer, pp 736-744.

37. Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, et al. (2018): Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb Cortex*. 28:3095-3114.

38. Craddock RC, James GA, Holtzheimer PE, 3rd, Hu XP, Mayberg HS (2012): A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum Brain Mapp*. 33:1914-1928.