DEEP EIGENVALUE-BASED PREDICTION OF FUNCTIONAL MAGNETIC RESONANCE IMAGING

A PREPRINT

Derek M. Berger*
Department of Computer Science
St. Francis Xavier University
Anigonish, Nova Scotia, Canada
dberger@stfx.ca

Jacob Levman
Canada Research Chair
St. Francis Xavier University
Anigonish, Nova Scotia, Canada
jlevman@stfx.ca

October 9, 2021

ABSTRACT

Resting-state functional magnetic resonance imaging (rs-fMRI) data has considerable potential for predicting neuropsychological and neurophysiological disorders, especially with modern machine learning (ML) and deep learning (DL) techniques. However, the low signal-to-noise ratio, and high dimensionality of this data make preparing rs-fMRI data for use in (ML) and/or (DL) challenging. We develop a preprocessing step for converting rs-fMRI images into spatially-rich 4D summary images by examining the eigenvalues of perturbations to the functional connectivity (FC). We use both classical ML algorithms and novel DL architectures developed to exploit this 4D information, and show that these "eigenperturbation" images have equal or greater predictive potential to the raw fMRI or other common FC-based approaches. We demonstrate the potential utility of these eigenperturbation images using the difficult ABIDE dataset, and obtain start of the art overall accuracies of [70%-80%].

Keywords fMRI · ABIDE · deep learing · convolutional neural network · random matrix theory · eigenvalues · perturbation

1 Introduction

Despite tremendous successes in computer vision and natural language processing, machine learning (ML) and deep learning (DL) algorithms capable of handling higher-dimensional natural data are much less well-established. The handling of 3D medical images is particularly challenging, due to necessary pre-processing, unique normalization challenges, larger compute costs, and limited sample sizes [see Singh et al., 2020, for a general review].

[insert tedious paragraph introducing fMRI here]

Functional magnetic resonance images (fMRI) are especially challenging. The 4D nature of fMRI means that a single preprocessed fMRI image can be hundreds of megabytes. A single fMRI is essentially a (3D) video file[2], and while there are well-known and well-tested networks that are freely-available pre-trained for processing 2D *images* (e.g. ResNet, ResNext, MobileNet), efficient and effective models for processing even 2D video data are still just beginning to be developed [Xie et al., 2018, Tran et al., 2018, Wang et al., 2020]. Attempts to process fMRI data in full are even less developed [cite examples].

In addition, there is only very limited publicly availabile fMRI data. While databases such as OpenNeuro [Markiewicz et al., 2021] exist and cover a wide variety of domains, the number of subjects and is typically far too small for machine learning applications, and the studies are of very low quality, and far too heterogenous to be usefully combined.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

[2]Except that while in a video file pixel values are in $[0, 255]$, in fMRI the voxel values can be any floating point value.

Perhaps the most popular publicly available fMRI dataset is the ABIDE I dataset, which includes 1112 resting-state fMRI (rs-fMRI) scans from 539 subjects with autism spectrum disorders (ASD) and 573 age-matched controls with typical development (TD) [Di Martino et al., 2014]. However, while the ABIDE I dataset is an appropriate size for ML and DL approaches, it is particularly challenging to work with due to the inclusion of subjects from 20 different sites. The resulting varied scanning parameters, demographics, and psychological characteristics means that state-of-the art classification accuracies are only on the order of 70-75% (e.g. Table 3).

Additionally, the ADHD-200 data [Bellec et al., 2017] is a pre-processed rs-fMRI dataset of 585 TD and 362 children with Attention Deficit Hyperactive Disorder (ADHD), also containing data from 8 sites and 17 different studies. As with ABIDE, accurate classification is highly challenging, and competition accuracies originally ranged only from 37.4–60.5% [Milham et al., 2012], with more recent results in the range of 69-76% [see Liu et al., 2021, for a brief review].

Thus, there is a need for *general* methods to extract useful features from rs-fMRI. Being able to handle such heterogenous and varied data is also essential for automated prediction procedures that are to be useful in clinical settings, or for any research that seeks to acquire truly generalizable insights from fMRI data.

1.1 Prediction from 4D Spatial or Spatiotemporal Data

1.1.1 Deep Learning Approaches

As should be clear from Table 3, all state-of-the-art classification accuracies are obtained either using a pure DL approach or a feature selection plus DL approach. While other hybrid approaches will likely be developed in the future and achieve even more impressive results, it is likely that any approach will incorporate *some* DL methods.

In deep learning, it is common to make a distinction between spatial and channel dimensions. For example, colour images, with three colour dimensions (red, green, and blue) and two spatial dimensions (height, width) are usually described as multi-channel 2D. Color video is thus multi-channel 3D: there are three color channels, and 3 spatial dimensions (height, width, and time). Likewise, MRI is one-channel 3D, diffusion MRI is multi-channel 3D, and rs-fMRI is one-channel 4D.

Channel dimensions are special in that their *ordering* is largely arbitrary: shuffling the order of the channels has no meaningful impact on the semantics or usable information in the inputs: channels are parallel (but not necessarily independent) features with identical spatial dimensions.

Given the success of the 2D convolutional neural network (CNN) in computer vision, the most natural approach to processing video is thus to take successful multi-channel 2D networks (e.g. ResNet) and expand the convolutions to be 3D. However, the computational costs of 3D convolution are much larger. While 2D convolution has a computational complexity of $\mathcal{O}(CHW)$ for channels C, height H and width W, 3D convolution has complexity $\mathcal{O}(CHWD)$ for depth D [Tran et al., 2019]. 2D CNNs typically crop or downsample images to e.g. $N_p = 256 \times 256 = 2^9$ pixels and have 3 channels in the input layers, and then internal layers have from 2^8 up to 2^{10} or more channels, with anywhere from tens to hundreds of such layers, i.e. the complexity is in practice more like over $\mathcal{O}(2^{17} * L)$ for a network with number of layers L. By contrast, even a low-resolution MRI scan with dimensions of $64 \times 64 \times 64$ is already 2^{18} voxels, and for fMRI, which may have, say, 120 or more timepoints, this is already approximately 2^{25} voxels with just a single channel. I.e. fMRI we expect to be two to three orders of magnitudes more costly to process than 2D images.

Similar rough calculations show that the leap from colour 2D to colour video would be only a one to two order of magnitude increase, and yet still much of the focus of existing implemeentations for 2D video processing focus on performance challenges [e.g. Tran et al., 2018, 2019, Wang et al., 2020].

fMRI is truly 4D (one-channel) spatio-temporal data: temporal information is *not* channel data, as the order of the volumes is (presumably) important in fMRI. To properly process this would require 4D convolution, which is not implemented in most popular frameworks, and would be too computationally expensive to use practically. We thus experiment with a number of techniques to reduce computational costs, as documented in the Methods.

1.2 Existing Approaches and Benchmarks

TODO: Verbally summarize other overall details of existing studies not already covered in Table 3. E.g. summarize:

- FC approaches based on ROI mean signal correlations
- graph-based approaches which further process the FC to extract graph metrics
- full approaches (no FC)

- embedding / autoencoder / learned representation approaches
- other exotic approaches (harmonization, formal modeling, ...)

1.3 Problems with Existing Approaches

Unfortunately, much of the research on the ABIDE dataset is of low-quality, and studies reporting accuracies above 71% all have non-trivial issues either in validation or the general fitting procedure.

1.3.1 Biased Feature Selection

Suppose we have a sample of data $\mathbf{X} \in \mathbb{R}^n$, targets $\mathbf{y}$, a feature selection (or engineering) procedure $S : \mathbb{R}^n \to \mathbb{R}^m$ parameterized by ω, and a model (e.g. classifier) f with hyperparameters θ, so that we can write $f(\boldsymbol{x}) = f_\theta(\mathbf{x})$ for $x \in S(\mathbf{X})$. Suppose we also have some validation procedure which operates on the outputs of f and produces some score or loss $\mathcal{L}(f_\theta(S(\mathbf{X})), \mathbf{y}) \in \mathbb{R}$ to be minimized, and computed via some validation procedure (e.g. mean accuracy from k-fold cross-validation). Then the loss is $\mathcal{L}_{\theta,\omega}(\mathbf{X}, \mathbf{y})$ or $\mathcal{L}(\theta, \omega)$, omitting the data. Finding an optimal set of features and hyperparameters thus requires finding optimal or good $\theta^\star$ and $\omega^\star$ values.

Clearly, feature selection and hyperparameter optimization can be combined conceptually and the whole process is identical to the usual optimization problem where we attempt to minimize $\mathcal{L}(\eta)$ for a function $g_\eta = f_\theta \circ S_\omega$ where $\eta = (\theta, \omega)$. Either ω and θ can be optimized sequentially, by choosing some starting $\hat{\theta}$, computing $\omega^\star = \arg\min_\omega \mathcal{L}(\hat{\theta}, \omega)$, and then computing $\theta^\star = \arg\min_\theta \mathcal{L}(\theta, \omega^\star)$ to get the optimal performance $\mathcal{L}(\theta^\star, \omega^\star)$, or in combination, by simply finding $\arg\min_\eta \mathcal{L}(\eta)$[3].

If $\mathcal{L}$ is also the final performance metric (e.g. the misclassification error), then to prevent biased performance (over)estimates, some kind of holdout or cross-validation procedure is needed. That is, we must partition $(\mathbf{X}, \mathbf{y})$ into mutually disjoint sets $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$, $(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$, and optionally $(\mathbf{X}_{\text{val}}, \mathbf{y}_{\text{val}})$, and ensure that $(\mathbf{X}_{\text{test}}$ and $\mathbf{y}_{\text{test}})$ are *never* used in any optimization procedure.

Feature selection is *not* exempt from this requirement. Using all data in feature selection means the reported result is $\mathcal{L}_{\eta^\star}(\mathbf{X}, \mathbf{y})$, where $\eta^\star = \arg\min_\eta \mathcal{L}_\eta(\mathbf{X}, \mathbf{y})$, which is a *biased* and *overfit* estimate of performance (regardless if $\mathcal{L}$ uses k-fold internally) because such a procedure uses *all* the data multiple times, and simply chooses the largest value. With this procedure, there is no way to know if feature selection or engineering has actually resulted in meaningful improvements compared to no feature selection/engineering.

In fact, as most DL is viewed as performing implicit feature engineering, choosing the "best" feature set using all data, and then reporting results only for that feature set would be the equivalent in deep learning of performing multiple model runs *and multiple evaluations on the test set*, and then just reporting the "best" test result. Thus, any study which performs biased feature selection cannot be meaningfully compared to a study that makes proper use of a fully held out test set, i.e. one which reports $\mathcal{L}_{\eta^\star}(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$, where $\eta^\star = \arg\min_\eta \mathcal{L}_\eta(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$ or $\eta^\star = \arg\min_\eta \mathcal{L}_\eta(\mathbf{X}_{\text{val}}, \mathbf{y}_{\text{val}})$

This "double-dipping" [Kriegeskorte et al., 2009] problem is well-known and not at all benign, and for data like MRI, can inflate accuracy estimates by tens of percentage points [Wen et al., 2019]. Unfortunately, many studies on ABIDE I reporting overall accuracies exceeding 71% either clearly involve such a biased feature-selection procedure, or provide insufficient details to rule out whether feature selection was biased in this manner (see Table 3, "Biased" column). We discuss how we avoid this problem in the Methods.

1.3.2 Heterogeneity and Small Validation Sets

For heterogeneous data such as ABIDE, with 20 different data sources, validation itself poses a unique challenge. We can see from Table 3 and from Table 9 of Sakai and Yamada [2019] that with the heterogeneous ABIDE data, the apparent validation accuracy is highest when the number of validation folds increases (or equivalently, as the average size of the validation set decreases). This was also demonstrated on ABIDE in Iidaka [2015], where accuracies of 77.2%, 86.9%, and 90.3% for 2-, 10-, and 50-fold respectively were obtained[4].

While this would not generally be expected from data sampled from a single distribution, when taking small samples from a heterogeneous mixture of distributions, this outcome is unsurprising. As the number of folds or validation sets, k, approaches 2, the likelihood that a validation sample comes from a distribution that was never seen even once during training increases, and as k approaches the number of samples, the reverse occurs.

[3] Of course the sequential optimization could be performed in the reverse order

[4] These accuracies are quite high likely due to the use of an age-restricted subset of ABIDE and biased feature selection (see section 1.3.1)

Given that ABIDE has about 20 different sites (XX from NYU and XX from ... being the majority), and that there are at most just over 1000 subjects in the full ABIDE dataset, the reported overall accuracy is largely a byproduct of the folding strategy and randomization or seed. Thus, assessment of randomization failure, or approaches like leave-one-site-out (LOSO) validation or stratified k-fold with stratification on site or scan parameters are *essential* to a proper assessment of model performance on the ABIDE dataset[5].

Without such safeguards, one could unwittingly simply stumble upon a seed which performs well simply due to randomization failure. Since most serious ABIDE studies have approaching 1000 samples, an average of just 1 more correctly classified subject per fold will result in an overall accuracy increase of 0.5% for 5-fold, and 1% for 10-fold. Since ABIDE contains data from 20 different sites, it is not hard to see that one could achieve apparently impressive gains in overall accuracy, even with k-fold, simply by "seed hacking" (perhaps quite unwittingly) until validation sets overall tend to contain subjects from already-seen distributions [Picard, 2021].

Of the studies reviewed in this paper, *none* except Ingalhalikar et al. [2021] and Byeon et al. [2020] mention or consider stratification or report fold site distributions. Thus it is unclear whether strangely impressive results like a 5-fold accuracy of over 75% in Yang et al. [2020] are the result of model design, validation procedures, or "lucky" splits.

Additionally, all the reported state-of-the-art overall accuracies are just within a few percentage points of each other. Even on datasets orders of magnitudes larger, differences in seeds can result accuracy differences this large [Picard, 2021]. Thus without detailed reporting of validation splits, or setting aside established test subjects, it is almost impossible to determine if apparent model improvements are actually meaningful. We also make efforts to improve this state of affairs.

2 Methods

An fMRI image $\boldsymbol{F}$ is a tensor or array with shape `(H,W,D,T)`, where `H`, `W`, `D`, and `T` are the sizes of the height, width, depth, and time dimensions, respectively. The flattened array F thus has shape $(N = \texttt{H} \times \texttt{W} \times \texttt{D}, \texttt{T})$, such that `N` is the total number of voxels in the image (air voxels included). A region of interest (ROI) we denote with R, and each of the n ROIs R_i of $\boldsymbol{F}$ has shape $(N_i, \texttt{T})$, where $\sum_i^n N_i = N$. ROIs are defined by an *atlas*, which is an array of integers of shape `(H,W,D)`, with integer values up to n.

2.1 ROI-Based Feature Extraction

By far the vast majority of studies (see Table 3) use atlas-based ROIs in order to convert the 4D fMRI image to a 2D array. Either ROI mean signals are used, as in El-Gazzar et al. [2019], or the (upper triangle) of the matrix of correlations (FC) between all such mean signals are used as input features.

The mean is just a statistic, i.e. a formal method of summarizing data. Other statistics could be just as useful here. In particular, rather than a measure of location, a measure of scale, like the variance, could be used to summarize the ROI signals. For example, we could compute a "variance signal" which, at each timepoint, is the standard deviation of the ROI voxels at that timepoint.

If ROI mean or variance signals are used directly as features, the input tensor has shape (n, T). If the correlations of these ROI signals are used as features, then since the correlation matrix is symmetric with 1s on the diagonals, then there are $n(n-1)/2$ unique features.

Finally, this correlation matrix can itself be summarized, either by graph theoretic measures as in Yin et al. [2021] and Mostafa et al. [2019], which compute the eigenvalues of the Laplacian of the thresholded correlation matrix, or simply by the eigenvalues.

Thus, we extract eight ROI-based features: the ROI mean signals (μ_{ROI}), the ROI standard deviation signals σ_{ROI}, the correlations between these $r_{\mu,\text{ROI}}$ and $r_{\sigma,\text{ROI}}$, the eigenvalues of the Laplacian of these, thresholded at T, $\lambda_{\mu,\text{ROI},T}$ and $\lambda_{\sigma,\text{ROI},T}$, and the direct eigenvalues of the correlations, $\lambda_{\mu,\text{ROI}}$ and $\lambda_{\sigma,\text{ROI}}$, as well as concatenations of these Their shapes are in 2.1.

As the highest quality existing results and studies use mostly either the CC-200 or CC-400 [cite] atlases for extracting ROI sequences, we test results for both of these atlases. We also compare concatenated pairings of mean and standard

[5] although mean accuracy on e.g. LOSO will also be inflated relative to 5- or 10-fold, as most sites have only a small number of samples, and there are almost 20 sites)

Feature	Name	Shape
ROI means	μ_{ROI}	(n, T)
ROI standard deviations (sds)	σ_{ROI}	(n, T)
ROI descriptive sequences	$(\mu_{\text{ROI}}, \sigma_{\text{ROI}})$	$(2, n, T)$ or $(2n, T)$
Correlation of ROI means	$\mathbf{r}_{\mu,\text{ROI}}$	(n, n)
Correlation of ROI sds	$\mathbf{r}_{\sigma,\text{ROI}}$	(n, n)
Correlation descriptives $[\mathbf{r}_{\mu,\text{ROI}}; \mathbf{r}_{\sigma,\text{ROI}}]$	$\mathbf{r}_{\text{ROI}}$	$(2, n, n)$ or $(2n, n)$
T-thresholded Laplacian eigenvalues of $\mathbf{r}_{\mu,\text{ROI}}$	$\lambda_{\mu,\text{ROI},T}$	$n-1$
T-thresholded Laplacian eigenvalues of $\mathbf{r}_{\sigma,\text{ROI}}$	$\lambda_{\sigma,\text{ROI},T}$	$n-1$
Concatenated T-thresholded Laplacian eigenvalues	$\lambda_{\text{ROI},T}$	$(2, n-1)$
Eigenvalues of $\mathbf{r}_{\mu,\text{ROI}}$	$\lambda_{\mu,\text{ROI}}$	$n-1$
Eigenvalues of $\mathbf{r}_{\sigma,\text{ROI}}$	$\lambda_{\sigma,\text{ROI}}$	$n-1$
Concatenated eigenvalues $[\mathbf{r}_{\sigma,\text{ROI}}; \mathbf{r}_{\sigma,\text{ROI}}]$	λ_{ROI}	$(2, n-1)$

Table 1: ROI-based feature shapes. n = number of ROIs. $[A; B]$ = concatenation of A and B along new initial (channel) dimension for deep learning, or along initial dimension, for classical ML.

T	N subjects
78	25
116	119
124	4
146	59
152	29
176	211
196	129
202	1
206	28
232	1
236	86
246	56
296	120
316	3

Table 2: Heterogeneity in ABIDE number of timepoints. Mean number of timepoints = 193, median = 176. Standard deviation = 58.1

2.1.1 Sampling Heterogeneity and Unequal Timepoints

Prior to the development of DL and CNNs, handling sequences of different lengths in predictive algorithms was a serious challenge. While sequences lengths that differ by multiple orders of magnitude are still a problem, smaller differences can generally be handled quite simply by padding the beginning of short sequences with zeroes [cite]. This seems to work especially well with CNNs [cite]. For certain types of data, long sequences can also be truncated, when appropriate. Given the open-ended and non-specific nature of the rs-fMRI "task" we believe truncation is (reasonably) harmless here.

For correlation-based methods, most previous studies ignore these issues, as the shape of the correlation matrix is a function of the number of ROIs, and not the number of timepoints. In the interest of comparison to these methods, we do the same, and use all timepoints to compute correlations.

However, for methods using the actual ROI sequences, or the eigenvalues, the number of timepoints effects the extracted feature length. The distribution of timepoints in the ABIDE data is shown in Table 2.1.1.

For simplicity, similarity to the size of the CC-200 atlas, and to keep extracted features comparable, we use only the first 200 timepoints for ROI sequences (μ_{ROI}, σ_{ROI}, and their concatenation), and use only the first 201 timepoints when extracting eigenvalues (as an $n \times n$ correlation matrix has at most $n-1$ unique eigenvalues. For subjects with less timepoints, the sequences are *first* normalized, and then front zero-padded to a length of 200. E.g. subjects with 78 timepoints have the 77 non-zero eigenvalues computed, normalized, and then 113 zeros are added to the front of the sequence.

TR (s)	N subjects
1.5	81
1.7	56
2	519
2.2	26
2.5	58
3	131

Table 3: Heterogeneity in ABIDE repetition times. Note many subjects have different TR values very close to 1.666 due to rounding issue, so are just reported here as 1.7 for simplicity

2.1.2 Eigenvalue Normalization

Both the graph Laplacian and correlation matrix eigenvalues are all positive, by virtue of the positive semi-definite nature of the parent matrices. Additionally, the size of the largest and smallest eigenvalues are dependent on the matrix dimensions. As per Random Matrix Theory, the spacings between eigenvalues tends to be highly exponential, e.g. the ith largest eigenvalue λ_i will be orders of magnitude larger than the smallest eigenvalue. In other words, the eigenvalues have a strongly increasing monotonic trend.

Such a trend can be a problem for deep learners using convolutions or that apply the same weights to each timepoint (e.g. LSTM). Thus, we potentially need to remove this trend for such methods [cite or argue].

Thus, non-linear normalization methods might be preferable to linear methods. We test two methods for normalizing eigenvalue-based features: MinMax Normalization (linear), where $\tilde{\lambda}_i = (\lambda_i - \lambda_{\text{min}})/(\lambda_{\text{max}} - \lambda_{\text{min}})$, and "exponential normalization", where we

2.2 Whole-Brain-Based Feature Extraction

[Describe eigenvalue perturbation procedure here]

2.2.1 Motivation for Eigenvalue Perturbation

Due to the multi-site nature of the ABIDE data, the number of timepoints and repetition time (TR) vary across scans. ([TODO]: see Table or Histogram). However, deep or machine learning algorithms require inputs to be of the same size (if not generally, then in mini-batches of more than one input). While cropping, padding or interpolation are usually used to overcome such issues, these techniques are more difficult to justify when the underlying sampling rates differ and signals are noisy, and are likely to lead to poor generalization and memorization when sampling rates and image sizes vary significantly. I.e. padded inputs occupy completely different regions of space and so will form a clearly separable cluster, cropped inputs with different sampling rates represent different amounts of measurement time, and interpolated inputs are likely to have increased correlations between features (voxels) unless the interpolation is downsampling and does not apply any averaging or smoothing (e.g. as in nearest-neighbour interpolation).

Thus there is a challenge of converting all images to some common representation which is relatively insensitive to these temporal differences, or which eliminates these differences while retaining rich information about the original signals.

One recent finding has been that the functional connectivity of rs-fMRI is surprisingly robust to the the underlying sampling rate [Huotari et al., 2019, Shakil et al., 2016]

2.3 Data

2.3.1 ABIDE Data

The ABIDE I dataset [Di Martino et al., 2014] is a publicly acessible rs-fMRI dataset of over 1112 subjects, 539 of which are diagnosed with Autism Spectrum Disorder (ASD), and 573 of which are typical development (TD). The data are collected from 17 different sites with varying scan parameters, and subjects vary considerably in age, making the dataset analytically challenging. However, the high heterogeneity of the data also makes it more suitable for testing the generalizability of modern predictive methods, and methods that perform well on the ABIDE data *a priori* have more potential clinical utility.

The ABIDE I data is available fully-preprocessed, and investigators can choose from a number of pre-processing pipelines and options. The main options involve filtering (bandpass filtering, global signal regression), and a choice of one of four pipelines [Di Martino et al., 2014]. However, to keep analyses consistent with similar papers Abraham et al. [2017], Mostafa et al. [2019], Yin et al. [2021], Heinsfeld et al. [2018] we use the subjects from the Configurable Pipeline for the Analysis of Connectomes [CPAC; Cameron et al., 2013], and exclude subjects that fail to pass quality control checks from three independent experts [see Abraham et al., 2017, for details], giving a final total of 871 subjects (XXX ASD, XXX TD).

2.3.2 ADHD-200

TODO: If things all work well with ABIDE, download this dataset and use our methods, and also compare to existing research there.

2.4 Models and Architectures

[**TODO**: Describe depthwise-separable Conv3d to ConvLSTM3d architecture here, and motivation.] [**TODO**: Describe SVM, RandomForest, AdaBoost models for comparison]

2.5 Fitting Procedure

[**TODO**: Describe automated hyperparameter tuning via Optuna]

3 Results

Study	Pipeline	Model	Features	Biased	Atlas or N_{ROI}	N	N_{ASD}	N_{TD}	Validation	OA
El-Gazzar et al.	CPAC	Conv1D	ROI means	N	HO	1100	?	?	5-fold	64.0
Yang et al.	CPAC	ANN	ROI-FC	N	CC-400	1035	505	530	5-fold	75.27^{a}
Almuqhim and Saeed	CPAC	AE+ANN	ROI-FC	N	CC-200	1035	505	530	10-fold	70.8
Eslami et al.	CPAC	AE+ANN	ROI-FC	N	CC-200	1035	505	530	10-fold	70.3
Sherkatghanad et al.	CPAC	CNN	ROI-FC	N	CC-400	1035	505	530	10-fold	70.2
Heinsfeld et al.	CPAC	ANN	ROI-FC	N	CC-200	1035	505	530	10-fold	70.0
Ingalhalikar et al.	DPARSF	ANN	ROI-FC+	?	CC-200	988	432	556	10-fold LOSO	71.4
Wang et al.	CPAC	GCN+ensemble	ROI-FC	?	multiple	949	419	530	10-fold	75.9 (70.7-72.5)[b]
Yin et al.	custom	AE+SVM	graph	N	264^{c}	871^{d}	403^{d}	468^{d}	holdout[d]	78.3^{d}
Shao et al.	CPAC	FS+GCN	ROI-FC	Y	HO	871	403	468	10-fold	79.5
Mostafa et al.	custom	FS+LDA	graph	?	264^{c}	871	403	468	10-fold	77.7
Parisot et al.	CPAC	FS+GCN	graph	?	HO	871	403	468	10-fold	70.4
Khosla et al.	CPAC	CNN	ROI-FC+	N	CC-200	$774/393^{e}$	$379/163^{e}$	$395/230^{e}$	holdout	72.8
Iidaka	custom	FS+PNN	ROI-FC	Y	AAL	640^{f}	312	328	10-fold	86.9
Byeon et al.	CPAC	CNN+RNN+	graph	N	BNA	575^{g}	270	305	5-fold[h]	74.5
Wang et al.	DPARSF	FS+SVM	ROI-FC	Y	35	531	255	276	LOSO	75.0-95.2
Li et al.	CPAC	ANN	ROI-FC	N	HO	370	186	184	5-fold	67.6% to 84.9%
Kazeminejad and Sotero	CPAC	FS+SVM	graph	N	AAL	342^{i}	?	?	10-fold	69.0

Table 4: Best overall accuracies of existing approaches using only over 200 subjects and rigorous validation strategies. Sorted by sample size and then accuracy. N = total number of subjects, N_{ASD} = total number of ASD subjects, N_{TD} = total number of TD subjects. DPARSF = [see Yan, 2010]. LOSO = Leave-one-site-out cross-validation. LOOCV = leave-one-out cross-validation. ANN = Aritificial Neural Network (linear layers). PNN = Probabilistic neural network [Specht, 1990]. CNN = Convolutional neural network RNN = Recurrent neural network GCN = graph convolutional network. AE = Autoencoder. RF = Random forest. SVM = support vector machine. LDA = Linear Discriminant Analysis. ROI-FC = ROI-based FC matrix ("+" indicates extra processing on ROI-FC). graph = graph-theoretic measures and metrics, usually extracted from ROI-FC. DA = domain adaptation CCS = Connectome Computation System [Xu et al., 2015]. CC-200/400 = Cameron-Craddock 200/400. HO = Harvard-Oxford Atlas. AAL = Automated anatomical labeling [Tzourio-Mazoyer et al., 2002]. BNA = BrainNetome Atlas [Fan et al., 2016] OA = Overall Accuracy. FS = feature selection. Biased = feature selection using full data.

[a] Note this paper also reports 5-fold accuracies of 71-72% with simple logistic regression, ridge regression, and linear SVM, which seems very difficult to square with other studies reviewed here, especially since 5-fold should be much more difficult than 10-fold, unless perhaps stratification is involved

[b] Bracketed values are values without ensembling.

[c] 264 ROI Atlas from [Power et al., 2011]

[d] Authors in fact use 80% of data for training and feature selection, and validate using 10-fold only on a final set held-out from feature selection of 174 subjects.

[e] Train/test. Training subjects from ABIDE I, testing from ABIDE II.

[f] Only subjects less than 20 years old were used for this study.

[g] Included only subjects with a TR of 2.0s

[h] Stratified for both ASD/TD and site ratios

[i] Only subjects between 10-15 years of age.

References

Satya P. Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás. 3D Deep Learning on Medical Images: A Review. *Sensors*, 20(18):5097, January 2020. doi:10.3390/s20185097.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. *arXiv:1712.04851 [cs]*, July 2018.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. *arXiv:1711.11248 [cs]*, April 2018.

Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video Modeling with Correlation Networks. *arXiv:1906.03349 [cs]*, May 2020.

Christopher J. Markiewicz, Krzysztof J. Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O. Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncalves, Anita Jwa, and Russell A. Poldrack. OpenNeuro: An open resource for sharing of neuroimaging data. Preprint, Neuroscience, June 2021.

A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, and M. P. Milham. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, June 2014. ISSN 1476-5578. doi:10.1038/mp.2013.78.

Pierre Bellec, Carlton Chu, François Chouinard-Decorte, Yassine Benhajali, Daniel S. Margulies, and R. Cameron Craddock. The Neuro Bureau ADHD-200 Preprocessed repository. *NeuroImage*, 144:275–286, January 2017. ISSN 1053-8119. doi:10.1016/j.neuroimage.2016.06.034.

Michael Milham, Damien Fair, Maarten Mennes, and Stewart Mostofsky. The adhd-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6: 62, 2012. ISSN 1662-5137. doi:10.3389/fnsys.2012.00062.

Shuaiqi Liu, Ling Zhao, Xu Wang, Qi Xin, Jie Zhao, David S. Guttery, and Yu-Dong Zhang. Deep Spatio-Temporal Representation and Ensemble Classification for Attention Deficit/Hyperactivity Disorder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1–10, 2021. ISSN 1558-0210. doi:10.1109/TNSRE.2020.3019063.

Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video Classification With Channel-Separated Convolutional Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.

Nikolaus Kriegeskorte, W. Kyle Simmons, Patrick S. F. Bellgowan, and Chris I. Baker. Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(5):535–540, May 2009. ISSN 1546-1726. doi:10.1038/nn.2303.

Junhao Wen, Jorge Samper-González, Alexandre M Routier, Simona Bottani, Stanley Durrleman, Ninon Burgos, and Olivier Colliot. Beware of feature selection bias! Example on Alzheimer's disease classification from diffusion MRI. In *2019 OHBM Annual Meeting - Organization for Human Brain Mapping*, Rome, Italy, June 2019.

Koji Sakai and Kei Yamada. Machine learning studies on major brain diseases: 5-year trends of 2014–2018. *Japanese Journal of Radiology*, 37(1):34–72, January 2019. ISSN 1867-108X. doi:10.1007/s11604-018-0794-4.

Tetsuya Iidaka. Resting state functional magnetic resonance imaging and neural network classified autism and control. *Cortex*, 63:55–67, February 2015. ISSN 0010-9452. doi:10.1016/j.cortex.2014.08.011.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv:2109.08203 [cs]*, September 2021.

Madhura Ingalhalikar, Sumeet Shinde, Arnav Karmarkar, Archit Rajan, D Rangaprakash, and Gopikrishna Deshpande. Functional connectivity-based prediction of Autism on site harmonized ABIDE dataset. *IEEE Transactions on Biomedical Engineering*, pages 1–1, 2021. ISSN 1558-2531. doi:10.1109/TBME.2021.3080259.

Kyoungseob Byeon, Junmo Kwon, Jisu Hong, and Hyunjin Park. Artificial Neural Network Inspired by Neuroimaging Connectivity: Application in Autism Spectrum Disorder. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 575–578, February 2020. doi:10.1109/BigComp48618.2020.00013.

Xin Yang, Paul T., and Ning Zhang. A Deep Neural Network Study of the ABIDE Repository on Autism Spectrum Classification. *International Journal of Advanced Computer Science and Applications*, 11(4), 2020. ISSN 21565570, 2158107X. doi:10.14569/IJACSA.2020.0110401.

Ahmed El-Gazzar, Mirjam Quaak, Leonardo Cerliani, Peter Bloem, Guido van Wingen, and Rajat Mani Thomas. A Hybrid 3DCNN and 3DC-LSTM Based Model for 4D Spatio-Temporal fMRI Data: An ABIDE Autism Classification Study. In Luping Zhou, Duygu Sarikaya, Seyed Mostafa Kia, Stefanie Speidel, Anand Malpani, Daniel Hashimoto, Mohamad Habes, Tommy Löfstedt, Kerstin Ritter, and Hongzhi Wang, editors, *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*, Lecture Notes in Computer Science, pages 95–102, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32695-1. doi:10.1007/978-3-030-32695-1_11.

Wutao Yin, Sakib Mostafa, and Fang-xiang Wu. Diagnosis of Autism Spectrum Disorder Based on Functional Brain Networks with Deep Learning. *Journal of Computational Biology*, 28(2):146–165, February 2021. doi:10.1089/cmb.2020.0252.

Sakib Mostafa, Lingkai Tang, and Fang-Xiang Wu. Diagnosis of Autism Spectrum Disorder Based on Eigenvalues of Brain Networks. *IEEE Access*, 7:128474–128486, 2019. ISSN 2169-3536. doi:10.1109/ACCESS.2019.2940198.

Niko Huotari, Lauri Raitamaa, Heta Helakari, Janne Kananen, Ville Raatikainen, Aleksi Rasila, Timo Tuovinen, Jussi Kantola, Viola Borchardt, Vesa J. Kiviniemi, and Vesa O. Korhonen. Sampling Rate Effects on Resting State fMRI Metrics. *Frontiers in Neuroscience*, 13:279, 2019. ISSN 1662-453X. doi:10.3389/fnins.2019.00279.

Sadia Shakil, Chin-Hui Lee, and Shella Dawn Keilholz. Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states. *NeuroImage*, 133:111–128, June 2016. ISSN 1053-8119. doi:10.1016/j.neuroimage.2016.02.074.

Alexandre Abraham, Michael P. Milham, Adriana Di Martino, R. Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*, 147:736–745, February 2017. ISSN 1053-8119. doi:10.1016/j.neuroimage.2016.10.045.

Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R. Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17: 16–23, January 2018. ISSN 2213-1582. doi:10.1016/j.nicl.2017.08.017.

Craddock Cameron, Sikka Sharad, Cheung Brian, Khanuja Ranjeet, Ghosh Satrajit, Yan Chaogan, Li Qingyang, Lurie Daniel, Vogelstein Joshua, Burns Randal, Colcombe Stanley, Mennes Maarten, Kelly Clare, Di Martino Adriana, Castellanos Francisco, and Milham Michael. Towards Automated Analysis of Connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC). *Frontiers in Neuroinformatics*, 7, 2013. ISSN 1662-5196. doi:10.3389/conf.fninf.2013.09.00042.

Fahad Almuqhim and Fahad Saeed. ASD-SAENet: A Sparse Autoencoder, and Deep-Neural Network Model for Detecting Autism Spectrum Disorder (ASD) Using fMRI Data. *Frontiers in Computational Neuroscience*, 15:27, 2021. ISSN 1662-5188. doi:10.3389/fncom.2021.654315.

Taban Eslami, Vahid Mirjalili, Alvis Fong, Angela R. Laird, and Fahad Saeed. ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data. *Frontiers in Neuroinformatics*, 13:70, 2019. ISSN 1662-5196. doi:10.3389/fninf.2019.00070.

Zeinab Sherkatghanad, Mohammadsadegh Akhondzadeh, Soorena Salari, Mariam Zomorodi-Moghadam, Moloud Abdar, U. Rajendra Acharya, Reza Khosrowabadi, and Vahid Salari. Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network. *Frontiers in Neuroscience*, 13:1325, 2020. ISSN 1662-453X. doi:10.3389/fnins.2019.01325.

Yufei Wang, Jin Liu, Yizhen Xiang, Jianxin Wang, Qingyong Chen, and Jing Chong. MAGE: Automatic diagnosis of autism spectrum disorders using multi-atlas graph convolutional networks and ensemble learning. *Neurocomputing*, July 2021. ISSN 0925-2312. doi:10.1016/j.neucom.2020.06.152.

Lizhen Shao, Cong Fu, Yang You, and Dongmei Fu. Classification of ASD based on fMRI data with deep learning. *Cognitive Neurodynamics*, May 2021. ISSN 1871-4099. doi:10.1007/s11571-021-09683-0.

Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease. *Medical Image Analysis*, 48:117–130, August 2018. ISSN 1361-8415. doi:10.1016/j.media.2018.06.001.

Meenakshi Khosla, Keith Jamison, Amy Kuceyeski, and Mert R. Sabuncu. Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. *NeuroImage*, 199:651–662, October 2019. ISSN 1053-8119. doi:10.1016/j.neuroimage.2019.06.012.

Canhua Wang, Zhiyong Xiao, and Jianhua Wu. Functional connectivity-based classification of autism and control using SVM-RFECV on rs-fMRI data. *Physica Medica*, 65:99–105, September 2019. ISSN 1120-1797. doi:10.1016/j.ejmp.2019.08.010.

Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H. Staib, Pamela Ventola, and James S. Duncan. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Medical Image Analysis*, 65:101765, October 2020. ISSN 1361-8415. doi:10.1016/j.media.2020.101765.

Amirali Kazeminejad and Roberto C. Sotero. Topological Properties of Resting-State fMRI Functional Networks Improve Machine Learning-Based Autism Classification. *Frontiers in Neuroscience*, 12:1018, 2019. ISSN 1662-453X. doi:10.3389/fnins.2018.01018.

Yan. DPARSF: A MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. *Frontiers in System Neuroscience*, 2010. ISSN 16625137. doi:10.3389/fnsys.2010.00013.

D.F. Specht. Probabilistic neural networks and the polynomial Adaline as complementary techniques for classification. *IEEE Transactions on Neural Networks*, 1(1):111–121, March 1990. ISSN 1941-0093. doi:10.1109/72.80210.

Ting Xu, Zhi Yang, Lili Jiang, Xiu-Xia Xing, and Xi-Nian Zuo. A Connectome Computation System for discovery science of brain. *Science Bulletin*, 60(1):86–95, January 2015. ISSN 20959273. doi:10.1007/s11434-014-0698-3.

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. *NeuroImage*, 15(1):273–289, January 2002. ISSN 1053-8119. doi:10.1006/nimg.2001.0978.

Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R. Laird, Peter T. Fox, Simon B. Eickhoff, Chunshui Yu, and Tianzi Jiang. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral Cortex (New York, N.Y.: 1991)*, 26(8): 3508–3526, August 2016. ISSN 1460-2199. doi:10.1093/cercor/bhw157.

Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, and Steven E Petersen. Functional network organization of the human brain. *Neuron*, 72(4):665–678, November 2011. ISSN 0896-6273. doi:10.1016/j.neuron.2011.09.006.