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 1, 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 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].

Functional magnetic resonance images (fMRI) is especially challenging. The 4D nature of the data 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., 2020a]. 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.

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

*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.

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. This results in highly varied demographic and psychological characteristics and scanning parameters.

This heterogeneity makes ML and DL approaches difficult (see section 1.1 below for a review). However, being able to handle such heterogenous and varied data is 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 Existing Approaches and Benchmarks

TODO: Summarize results from notes in sections 3 and 4 here, both verbally listing various approaches.

1.2 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 clear, extremely serious issues either in validation or the general fitting procedure.

1.2.1 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 $S : \mathbb{R}^n \to \mathbb{R}^m$, 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 V 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). If we also say that S is parameterized by ω, such that we can write $S = S_\omega$, and we can also write that the loss is $\mathcal{L}_{\theta,\omega}(\mathbf{X}, \mathbf{y})$ or even just $\mathcal{L}(\theta, \omega)$. 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:

1. choose a sensible starting value $\hat{\theta}$ for θ [or instead start with $\hat{\omega}$]
2. compute $\omega^\star = \arg\min_\omega \mathcal{L}(\hat{\theta}, \omega)$ [or instead $\theta^\star = \arg\min_\theta \mathcal{L}(\theta, \hat{\omega})$]
3. compute $\theta^\star = \arg\min_\theta \mathcal{L}(\theta, \omega^\star)$ [or instead $\omega^\star = \arg\min_\omega \mathcal{L}(\theta^\star, \omega)$]
4. optimal performance is then $\mathcal{L}(\theta^\star, \omega^\star)$

or combined, by simply finding $\arg\min_\eta \mathcal{L}(\eta)$.

If our loss function $\mathcal{L}$ is also our final performance metric (e.g. the misclassification error), then if we do not want the above procedure to overfit our data and/or produce biased performance (over)estimates, we must employ some kind of holdout or cross-validation procedure. 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.

As per above, there is nothing special about feature selection. Feature selection + tuning is just another more complicated single fitting procedure, and so feature selection isn't *exempt* from the requirements of cross-validation or holdout. That is, in the notation of above, and putting back in the data used for the loss calculations, then

$$\mathcal{L}_{\eta^\star}(\mathbf{X}, \mathbf{y}), \quad \text{where} \quad \eta^\star = \arg\min_\eta \mathcal{L}_\eta(\mathbf{X}, \mathbf{y})$$

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. In particular, with the above 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 and 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 test set

Instead, one must report

$$\mathcal{L}_{\eta^\star}(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}}),$$

where either

$$\eta^\star = \arg\min_\eta \mathcal{L}_\eta(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) \quad \text{or} \quad \eta^\star = \arg\min_\eta \mathcal{L}_\eta(\mathbf{X}_{\text{val}}, \mathbf{y}_{\text{val}})$$

Unfortunately, almost all studies using a large number of subjects and reporting overall accuracies exceeding 71% either involve such a questionable feature-selection procedure, or provide insufficient details to rule out whether feature selection used all subjects (see 1.2.2, "Biased" column).

1.2.2 Heterogeneity and Validation

For heterogeneous data such as ABIDE, with 20 different data sources, validation itself poses a unique challenge. As is clear from Table 1.2.2, smaller validation set sizes (e.g. LOOCV, 10-fold) result in dramatically higher reported accuracies.

We believe this is an obvious and direct consequence of the site-level heterogeneity present in the ABIDE data.

Lack of Cross-Validation This is also clear from other reviews [Sakai and Yamada, 2019, Table 9]. Just as an example, Sakai and Yamada [2019] review 12 studies using the ABIDE data. Only four of these studies use more than 20% of the ABIDE data, and of these, only Heinsfeld et al. [2018] use rigorous k-fold validation. The others use out-of-bag or LOOCV validation, which have limited utility and are also well known to severely overestimate performance [cite].

Small Validation Sets Suppose, for any study, the total number of samples (e.g. subjects) is N. Then leave-one-out cross-validation (LOOCV) is N-fold cross validation, and validation using k folds is k-fold validation. We can see from 1.2.2 and from Table 9 of Sakai and Yamada [2019] that with the heterogeneous ABIDE data, the apparent validation accuracy is highest when k approaches N. Likewise, Iidaka [2015] report accuracies of 77.2%, 86.9%, and 90.3% for 2-, 10-, and 50-fold respectively[3].

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 k approaches 1, the likelihood that a validation sample comes from a distribution that was never seen even once during training increases, and as k approaches N, the reverse occurs. 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 (i.e. the extent to which the validation sets contain sites / TRs / number of timepoints not seen in training). Thus, assessment of randomization failure, or approaches like leave-one-site-out 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.

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 about 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. The only ways to make different studies properly comparable are

- stratified k-fold
- leave-one-site-out validation [Ingalhalikar et al., 2021]
- k-fold with small k

Of the studies reviewed in this paper, *none* except Ingalhalikar et al. [2021] mention stratification, seeds, or provide code for analyses. All the reported state-of-the-art overall accuracies are just within a few percentage points of each other. At about 900-1000 subjects typically for the ABIDE data, and with 5- or 10-fold cross-validation, this means that a difference in mean overall accuracy of just 1% can be due to just 2 mis-predicted subjects per fold. That is, one could likely very easily achieve difference, on average, of (i.e. due to a handful or less of) exceptional results could just as easily be the result of a "lucky" split (torch.manual seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision)

[3]These accuracies are quite high likely due to the use of an age-restricted subset of ABIDE and contamination / overfitting in the feature selection procedure, which used all subjects

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 1: 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.

[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.

Contamination A number of studies on the ABIDE data reporting overall accuracies above 71% use feature selection procedures which are given access to the full number of subjects. That is, given N subjects total, feature selection procedures are run on all N subjects, and then final validation is done with e.g. k-fold on the same N subjects, sometimes even using the exact same model that was used during feature selection [citations]. This procedure is not valid.

1.3 Prediction from 4D Spatial or Spatiotemporal Data

In deep learning, it is common to divide the dimensions of the data into spatial and channel dimensions. For example, colour images are usually described as multi-channel 2D, as there are typically three colour dimensions (RGB), and two spatial dimensions (height and width). 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. Intuitively, channels represent parallel features with the same spatial dimensions.

Thus color video is multi-channel 3D: there are three color channels, and 3 spatial dimensions (height, width, and time). 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 considerably larger (a factor of [cite] Tran et al. [2018]), and so much more successful papers have instead used "channel-separated" or (2 + 1)D convolutions (Tran et al. [2018]), or other methods like ..., or even methods which rethink how temporal information can be encoded via correlations Wang et al. [2020a].

fMRI is truly 4D (one-channel) spatio-temporal data. To properly process this would require 4D convolution, which is not implemented in most popular frameworks, and is extremely compute intensive. Simple three-channel, 2D video (i.e. three-channel 3D) already runs into compute problems, the 3D video data that comprise an fMRI image is especially difficult to handle effectively.

In particular, the increased GPU memory and compute demands means that hyperparameter tuning, an absolutely necessary step when investigating new network architectures, can be extremely expensive

Likewise, performance on messy data or data that has been corrupted even slightly results in often dramatically degraded performance [Metz et al., 2019, Dodge and Karam, 2017, Hendrycks and Dietterich, 2019, Azulay and Weiss, 2019, Rosenfeld et al., 2018]. Whether or nor fMRI images contain a large amount of "noise" relative to "signal" depends on one's analytic goals, however a number of factors (eye movement, heart rate, head movements, respiration ...) are detectable in fMRI signals [cite], and /empha priori not of clear causal relevance to a number of predictive goals.

1.4 Feature Extraction via Correlation Eigenvalues

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.

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]

1.5 Previous Approaches

TODO: Summarize various studies listed in notes below (3 and 4) in table or other condensed form.

2 Methods

2.1 ABIDE Data

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

The ABIDE I dataset is an appropriate size for ML and DL approaches, but is particularly challenging due to the inclusion of subjects from 20 different sites. This results in highly varied demographic and psychological characteristics and scanning parameters. While this heterogeneity makes ML and DL approaches difficult, it also means that if such an approach performs well on this data, that it has much more potential clinical utility and relevance.

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).

2.1.1 ROI Means

These are available from the CPAC pipeline for direct download [cite]. In the interest of reproducibility we simply use these directly.

2.1.2 ROI Standard Deviations

3 Key Articles

Note for below I also seem to use the CPAC minimal pipeline, I think.

Of studies that get above e.g. 70.3% accuracy, there is *almost always* some cheaty feature selection using all subjects, completely negating the final 10-fold or whatever accuracy values. No *honest* techniques make it above 70% accuracy across sites yet *except* Mostafa et al. [2019], Yin et al. [2021], who use eigenvalues of the graph of the Laplacian of the thresholded correlation matrix.

3.1 Use of Eigenvalues for Prediction! Mostafa et al. [2019]

3.2 Heinsfeld et al. [2018]

- 70% with C-PAC preprocessing pipeline
- used CC200 atlas to reduce to ROI feature vectors (i.e. not CNN)
- actually used functional connectivity as the predictor
- used 10-fold validation across all sites (good!)
- used a deep denoising autoencoder (really just MLP with some augmentation / dropout on inputs)

3.3 Eslami et al. [2019]

Decent. Very similar to ours in spirit.

- 1112 subjects from almost all sites
- CPAC pipeline with CC200 functional ROIs
- use only some correlations (e.g. select smallest and largest correlations (most negative?))
- train single-layer autoencoder first to do e.g. a non-linear PCA first for reduction / embedding
- use SMOTE to **augment their data!** (basically, linearly interpolate a new sample from 5 nearest samples), effectively doubles the training set size
- get to 70.3% accuracy (their augmentation adds only 1%)

3.4 Yin et al. [2021]

VERY GOOD, same authors as Mostafa et al. [2019].

- 871 subjects, custom preprocessing (minimal), 264 ROI parcellation, Pearson corrs
- also threshold map to largest (in abs. value) correlations (e.g. all correlations greater in absolute value then a threshold go to 1 or -1, so matrix becomes an adjacency matrix)
- then compute a Laplacian matrix (very simple, see article) from adjacency matrix
- then get eigenvalues from this
- compute a bunch of other graph-theoretic metrics from this Laplacian matrix
- interestingly **min-max normalize** their eigenvalues within subjects
- do not contaminate their autoencoder (proper splitting of training and testing)
  - however the splitting is a little bit fucky since it means pre-training
- without autoencoder pre-training and just using their graph features, get DNN acc of 76.2%
  - interestingly find that need a lower threshold (allow all correlations above 0.2) for this deep learner, which might be relevant for me since right now I am taking only top 25 eigenvalues, which is probably not enough (see Table 2 of their study, very interesting!)

Table 2: El-Gazzar et al. [2019] 5-fold cross-site results

Model	Accuracy	F1-score
AE MLP [8]	0.63 ± 0.02	0.64
SVM [5]	0.58 ± 0.04	0.60
1D Conv [7]	0.64 ± 0.06	0.64
3DCNN 1D (ours)	0.54 ± 0.02	0.50
3DCNN C-LSTM (ours)	0.58 ± 0.03	0.53

3.5 Khosla et al. [2019]

Note that choice of atlas doesn't matter, we could use this as a citation for motivating our non-atlas-based approach! Also this is a clever / good study that combines a lot of tricks and does seem to manage to get to 72.8%.

- CPAC pipeline, but did some extra expert / manual scrubbing and quality control (manual and automatic), exclude very young and very old, which brought them down to **163 ASD / 230 CTRL**
- looked at all ROIs from all atlases
- get to about 72.8% acc with CC200 parcellation and 3D-CNN (Table 2)
- various other models all get to around 71.2%, 71.7%, 72.3%, etc
- also generate a novel / clever new multi-channel 3D image where each channel is some different feature extracted from the connectivity info (e.g. channel 1 is voxel time series correlation with mean global time series, channel 2 is some other..., etc)

4 Other Articles

4.1 Li et al. [2020]

- 70% with C-PAC preprocessing pipeline
- similar number of subjects to us
- used sliding windows (32 timepoints long) of mean ROI sequences (HO Atlas, 111 ROIs)
- interesting comparison points is *sex classification accuracy* which was only between
- report no overall accuracy, but NYU hardest to classify (67.6% at best), USM easiest (up to 84.9%)
- validation splits are unclear

4.2 El-Gazzar et al. [2019]

- used Conv3D to Conv1D or ConvLSTM (they test both)
- CPAC pipeline for ABIDE
- selected for and cropped to 100 timepoints for the ABIDE-I dataset
- only use single-site validation (losers) also NYU and UM (easiest)
- also do one multi-site with 19 sites 1100 subjects
- patch-based training where prediction is average prediction over crops
- max 5-fold acc of 0.77 ± 0.05 on NYU with Conv3dConvLSTM3d

4.3 Shao et al. [2021]

- GCN (graph convolutional network) plus another network that learns feature weights
- use CPAC preprocessed data, but **discard 241 ghosty / bad-looking data** to get a total of 871 subjects
- used mean series on some of HO (Harvard-Oxford) atlas, standardized to zero mean and unit variance
- 111 ROIs total
- do feature selection and also test various models
- GCN gets $79.5 \pm 3.3\%$ acc (3.3 is sd), 10-fold validation

4.4 Dekhil et al. [2018]

- fMRI data for 123 ASD and 160 TD children and adolescents (for a total number of 283 subjects) from the National Database for Autism Research (NDAR: http://ndar.nih.gov)
- all Ps have both MRI and rs-fMRI scans
- do register, BET, slicetime, motion, and then spatial Gaussian spatial smoothing
- extract 34 ROIs via ICA
- features for prediction are the **power spectral densities**
- get very high accuracy, but isn't ABIDE, and is one site

4.5 Parisot et al. [2018]

- "Our analysis shows that our novel framework can improve over state-of-the-art results on both databases, with 70.4% classification accuracy for ABIDE"
- 871 subjects (remove a lot of bad scans) from CPAC pipeline
- GCN on correlation matrices between HO atlas mean ROI signals (plus some fancy normalization / filtering), with recursive feature selection
- stratified grouped 10-fold to get the 70.4% acc

4.6 Sakai and Yamada [2019]

See Table 9 of this one for ABIDE accuracies. There are some spuriously high and extremely unlikely out-of-bag accuracies of 0.90 or on tiny sample. Otherwise, of note is Iidaka [2015] who apparently get an LOOCV of 0.90 (??? dubious).

4.7 Iidaka [2015]

- only subjects under 20
- 640 subjects total
- 2-fold acc was 77.2%, 10-fold was 86.9%
- did do some bandpass filtering as pre-processing, drop first 5 volumes
- features are 90 AAL mean normalized ROI regions correlation matrix (r-values fisher normalized to Z-scores)

4.8 Li et al. [2018]

- pre-train stacked autoencoder on different data healthy rs-fMRI first
- ABIDE used Connectome Computation System preprocessed
- **regressed out height, age, sex, site** from among ROI correlations
- never get above 70.4% acc, and only within USM, so not good

4.9 Kazeminejad and Sotero [2019]

- CPAC pipeline, 116 AAL regions
- features are multiple correlation matrices (e.g. Pearson, Spearman, partial correlation, mutual information) converted to graphs via thresholding
- Gaussian SVM classifier (sklearn), step-up feature selection
- validation is 10-fold, but very dubious procedure, sounds / unclear how feature selection is being done here, almost certainly overfitting a lot
- no overall reported accuracy, just report stratified by ages, sad, lame, (lol Frontiers)

4.10 Khosla et al. [2019]

Note that choice of atlas doesn't matter, we could use this as a citation for motivating our non-atlas-based approach! Also this is a clever / good study that combines a lot of tricks and does seem to manage to get to 72.8%.

- CPAC pipeline, but did some extra expert / manual scrubbing and quality control (manual and automatic), exclude very young and very old, which brought them down to **163 ASD / 230 CTRL**
- looked at all ROIs from all atlases
- get to about 72.8% acc with CC200 parcellation and 3D-CNN (Table 2)
- various other models all get to around 71.2%, 71.7%, 72.3%, etc
- also generate a novel / clever new multi-channel 3D image where each channel is some different feature extracted from the connectivity info (e.g. channel 1 is voxel time series correlation with mean global time series, channel 2 is some other..., etc)

4.11 Wang et al. [2020b]

- 468 subjects, CPAC pipeline, AAL 116 mean time series
- don't really report clean overall accuracies because they are doing domain adaptation (training on non-NYU data and predicting NYU data) but basically when having similar validation sizes as us (e.g. about 40%) are also in the 69-73% range of accuracy

4.12 Sherkatghanad et al. [2020]

Would be good to try to replicate these dubious results

- CPAC pipeline, 871 "quality" images, CC400 parcellation
- shallow (but wide) CNN used directly on correlation matrix
- 10-fold acc of 70.2%

4.13 Wang et al. [2019]

Very curious, given how shallow network is and how simple approach is...

- reports a crazy cross-site accuracy of **over 90%**
- 531 subjects(255AD / 276CTRL), DPARSF pipeline, also did some QC
- 75-95% acc based on "leave-one-site-out"
- 90.6% acc on all sites with SVM
- key thing is a recursive, SVM-based step-down feature selection procedure on ROIs
- **HOWEVER is overfit because all 531 subjects used in feature selection... lol**

4.14 Yang et al. [2020]

Probably **worth trying to replicate this** given the stupid simplicity

- CC400 atlas, ROI correlation matrix, pathetically simply MLP, 1035 subjects
- claim to get 5-fold acc of 75.3%
- cross-validated grid-search of classifiers, so perhaps just overfitting the extra 4%?

4.15 Wang et al. [2021]

- 949 subjects, CPAC pipeline
- 75.86% accuracy, 10-fold
- multi-atlas Graph CNNs (ensemble learning), recursive feature elimination
- show single-atlas based models get 70-72% acc
- **feature selection subjects completely unclear, another cheat**

4.16 Ingalhalikar et al. [2021]

- most subjects, DPARSF pipeline, "cross-site harmomnization" with denoising autoencoder
- 71.35% 10-fold ACC with ANN

4.17 Yang et al. [2021]

77.74% 1-fold mean acc but with small number of subjects, only NYU (lame)

4.18 Almuqhim and Saeed [2021]

- claims Eslami et al. [2019] is current state of art (70.3%)
- CPAC pipeline, CC200, correlation matrix, only one triangle, 1/4 largest, 1/4 smallest
- sparse autoencoder to force embedding
- get 10-fold overall acc of 70.8%

4.19 Byeon et al. [2020]

- CPAC, only subjects with TR=2.0 (270 ASD, 305 TD) and good QC
- BrainNetome atlas mean signals (246 ROI)
- use first 146 timepoints only to make subjects all same size
- 74.5% accuracy, 5-fold

5 Headings: first level

See Section 5.

5.1 Headings: second level

Paragraph A footnote [4].

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 2020a.

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.

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.

[4] Sample of the first footnote.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Luke Metz, Niru Maheswaranathan, Jonathon Shlens, Jascha Sohl-Dickstein, and Ekin D. Cubuk. Using learned optimizers to make models robust to input noise. *arXiv:1906.03367 [cs, stat]*, June 2019.

Samuel Dodge and Lina Karam. A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–7, July 2017. doi:10.1109/ICCCN.2017.8038465.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Surface Variations. *arXiv:1807.01697 [cs, stat]*, April 2019.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv:1805.12177 [cs]*, December 2019.

Amir Rosenfeld, Richard Zemel, and John K. Tsotsos. The Elephant in the Room. *arXiv:1808.03305 [cs]*, August 2018.

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.

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.

Omar Dekhil, Hassan Hajjdiab, Ahmed Shalaby, Mohamed T. Ali, Babajide Ayinde, Andy Switala, Aliaa Elshamekh, Mohamed Ghazal, Robert Keynton, Gregory Barnes, and Ayman El-Baz. Using resting state functional MRI to build a personalized autism diagnosis system. *PLOS ONE*, 13(10):e0206351, October 2018. ISSN 1932-6203. doi:10.1371/journal.pone.0206351.

Hailong Li, Nehal A. Parikh, and Lili He. A Novel Transfer Learning Approach to Enhance Deep Neural Network Classification of Brain Functional Connectomes. *Frontiers in Neuroscience*, 12:491, 2018. ISSN 1662-453X. doi:10.3389/fnins.2018.00491.

Mingliang Wang, Daoqiang Zhang, Jiashuang Huang, Pew-Thian Yap, Dinggang Shen, and Mingxia Liu. Identifying Autism Spectrum Disorder With Multi-Site fMRI via Low-Rank Domain Adaptation. *IEEE Transactions on Medical Imaging*, 39(3):644–655, March 2020b. ISSN 1558-254X. doi:10.1109/TMI.2019.2933160.

Ming Yang, Menglin Cao, Yuhao Chen, Yanni Chen, Geng Fan, Chenxi Li, Jue Wang, and Tian Liu. Large-Scale Brain Functional Network Integration for Discrimination of Autism Using a 3-D Deep Learning Model. *Frontiers in Human Neuroscience*, 15:277, 2021. ISSN 1662-5161. doi:10.3389/fnhum.2021.687288.