# Demographic-Guided Attention in Recurrent Neural Networks for Modeling Neuropathophysiological Heterogeneity

Nicha C. Dvornek[1,2], Xiaoxiao Li[2], Juntang Zhuang[2], Pamela Ventola[3], and James S. Duncan[1,2,4,5]

[1]Radiology & Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA
[2]Biomedical Engineering, Yale University, New Haven, CT, USA
[3]Child Study Center, Yale School of Medicine, New Haven, CT, USA
[4]Electrical Engineering, Yale University, New Haven, CT, USA
[5]Statistics and Data Science, Yale University, New Haven, CT, USA

**Abstract.** Heterogeneous presentation of a neurological disorder suggests potential differences in the underlying pathophysiological changes that occur in the brain. We propose to model heterogeneous patterns of functional network differences using a demographic-guided attention (DGA) mechanism for recurrent neural network models for prediction from functional magnetic resonance imaging (fMRI) time-series data. The context computed from the DGA head is used to help focus on the appropriate functional networks based on individual demographic information. We demonstrate improved classification on 3 subsets of the ABIDE I dataset used in published studies that have previously produced state-of-the-art results, evaluating performance under a leave-one-site-out cross-validation framework for better generalizeability to new data. Finally, we provide examples of interpreting functional network differences based on individual demographic variables.

## 1 Introduction

Functional magnetic resonance imaging (fMRI) has begun to play a large role in characterizing the neuropathophysiology of psychiatric disorders. One example is in the characterization of autism spectrum disorder (ASD), a neurodevelopmental disorder that affects communication and behavior. ASD is extremely heterogeneous, presenting with a wide range of symptoms and severity of impairments. Early fMRI studies investigated small datasets with imposed homogeneity, e.g., restricting to one gender, age group, or level of functioning. However, this resulted in smaller datasets, largely irreproducible results and lack of generalization to new datasets. More recently, the popular large public Autism Brain Imaging Data Exchange (ABIDE) I resting-state fMRI dataset [4] has undergone extensive analysis, including the application of machine learning to classify ASD and healthy controls (HC) for the purpose of discovering neuroimaging biomarkers of ASD. However, even with the large amount of neuroimaging data, achieving high classification performance has been a challenge, likely due in part

to both the heterogeneity of the sample populations of each imaging site and the heterogeneity of the underlying neurological mechanisms of the disorder itself. Evidence for these potential reasons includes the much poorer performance of leave-one-site-out cross-validation (LOSO CV) compared to intrasite k-fold cross-validation [1,9].

One approach to mitigating the heterogeneity is to incorporate demographic information into the classification problem. Here, we refer to demographic variables as non-imaging, scalar variables that are often measured and easy to obtain, such as gender, age, or IQ. Demographic information can be incorporated in different ways depending on the classification model. For example, the demographic information can be fused at different layers in a standard feedforward neural network [7,12] or used as targets for prediction [7]. Furthermore, demographic information can be combined in model specific ways, e.g., to define the edges in graph-based models [13] or to set the initial state of recurrent neural network models [11,6]. However, none of these approaches aim to modulate the underlying neurological differences that may be describing the heterogeneity in ASD.

To model disorder heterogeneity in terms of changes in the underlying functional processing, we propose a demographic-guided attention module to enhance a recurrent neural network model for processing fMRI time-series data. While the attention scores are computed across time, we can interpret the resulting context as guiding attention to different functional networks. In addition to using the demographic information to help identify which functional networks to attend to in classifying ASD or HC, we propose a novel loss for computing more diverse queries for each attention head to better model the sample heterogeneity. We compare our proposed methods to other ways of handling demographic data on 3 subsets of the ABIDE dataset, matched to previous studies that have previously demonstrated state-of-the-art results from the fMRI data alone. We achieve some of the highest accuracy of ABIDE classification under LOSO CV. Finally, we give examples of functional networks that may undergo diverse processing in ASD based on individual demographic factors.

## 2 Methods

We build on recent models for predicting from fMRI time-series data that use recurrent neural networks with long short-term memory (LSTM). To model the heterogeneity of ASD, we apply a generalized attention mechanism that is guided by individual demographic characteristics. The context learned from the attention mechanism is then used to bias the LSTM outputs, allowing the model to focus on different functional networks based on individual non-imaging characteristics (Fig. 1).

### 2.1 Network Architecture

**Baseline LSTM for fMRI Time-Series** The baseline LSTM network to predict from fMRI time-series was first proposed by Dvornek et al. [5]. The fMRI

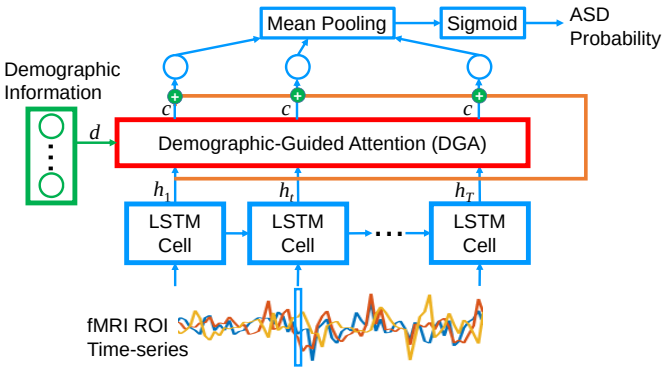



Fig. 1: Demographic-guided attention network for classification of ASD/HC from fMRI.

time-series with length $T$ from regions of interest (ROIs) in a predefined brain parcellation is first input to the LSTM layer. Then the output of the LSTM cell at each timepoint $h_t \in R^n$ is input to a fully connected (FC) layer with weights shared across time. The outputs of the FC layer are averaged across time and input to a sigmoid activation function to produce the probability of ASD label.

**Demographic-Guided Attention** We propose to incorporate functional network differences resulting from disease heterogeneity through a generalized attention mechanism [14]. The attention mechanism can be described as a function mapping a query and key-value pair to some output, often referred to as the context or a head. In our work, the query is defined by the demographic information, and the key and value are defined by the outputs of the LSTM layer $h_t$. Applying the scaled dot product attention [14], the context vector $c$ is computed by

$$c = att\left(d, \{h_t\}\right) = \sum_{t=1}^{T} softmax\left[{(W_q d)^T (W_k h_t)}/{\sqrt{m}}\right] W_v h_t, \tag{1}$$

where $d \in R^l$ is the vector of demographic information; $softmax\left[a_t\right] = {\exp(a_t)}/{\sum_{j=1}^{T} \exp(a_j)}$ with $a_t = {(W_q d)^T (W_k h_t)}/{\sqrt{m}}$; and $W_q \in R^{m \times l}$, $W_k \in R^{m \times n}$, and $W_v \in R^{m \times n}$ are weight matrices that operate on $d$ or $h_t$ to define the query, key, and value vectors, respectively. In this work, we set $m = n$.

**Residual Connection for Modeling Heterogeneity** In standard attention approaches, the context vector is concatenated with the LSTM output [2] or the context vectors alone [14] are used as input to the following layers. Here, we propose to use the context to bias the output of the LSTM layer, changing the focus on LSTM nodes that should be emphasized based on the demographic information. We do this by simply adding a residual connection betwen the output of each attention head $k$ and the output of the LSTM layer, $c_k + h_t$ (Fig.

1, orange path). The summed outputs are then processed in a similar way as the baseline LSTM model, using a FC layer with shared weights and averaging the FC outputs over time. If multiple attention heads are used, then each head is separately processed with a different FC layer, and the maximum score across the heads is passed to the sigmoid layer to represent probability of ASD. The rationale for keeping the maximum score is that only one mode of functional network patterns may be indicative of ASD.

### 2.2 Query Diversity Loss

A single attention module allows for attending to different LSTM nodes based on the demographic information. However, this assumes then that two individuals with the same demographic profile must share the same underlying neuropathology. To allow for even greater diversity in modeling disease heterogeneity, we can include more attention heads that will learn different contexts. To encourage the different attention heads to capture different underlying neuropathological modes, we propose the following query diversity loss (QDL):

$$L_{QD} = \sum_{i=1}^{N} \sum_{j=1}^{K-1} \sum_{k=j+1}^{K} \left| \frac{q_{ij}^T q_{ik}}{\|q_{ij}\| \, \|q_{ik}\|} \right| \tag{2}$$

where $N$ is the number of subjects and $q_{ij} = W_{q_j} d_i$ is the $n$-dimensional query vector for attention head $j$. QDL computes the cosine proximity for all query vectors $q_{ij}$ for subject $i$. Minimizing QDL thus encourages projection of the demographic information into orthogonal subspaces, which capture complementary information, before comparing to the keys to compute the attention scores.

The total loss $L$ is then

$$L = L_C + \lambda L_{QD}, \tag{3}$$

where $L_C$ is the classification loss (e.g., binary cross-entropy) and $\lambda$ is a hyperparameter controlling the contribution of QDL.

### 2.3 Interpretation of Attention as Neuropathological Heterogeneity

We first interpret each node of the LSTM as modeling a functional network. While different attribution methods can be applied, we follow Dvornek et al. and assign ROIs to a network if the LSTM weights for the ROI inputs have large magnitude ($> 3$ standard deviations above mean weight magnitude) [5].

The proposed model uses the context computed by the demographic-guided attention module as a bias for the LSTM outputs. Since each node of the LSTM is interpreted as processing the signal corresponding to some functional network, we interpret the demographic information as providing context for deciding which functional networks should be given more attention in performing ASD classification, i.e., we measure the demographic-guided attention to a functional network $f$ as $c(f)$. We then assess the coupling between a functional network and

a demographic variable by computing the correlation between the demographic variable $d(i)$ and the context $c(f)$ for functional network $f$ across subjects. Different patterns of attention for a functional network in different attention heads allows for modeling greater neuropathological heterogeneity.

## 3 Experiments

### 3.1 Data

We use resting-state fMRI data from the multisite ABIDE I dataset [4] which was released by the Preprocessed Connectomes Project [3]. To demonstrate robustness of our approach and directly compare with results from the literature, we analyzed the same subsets of data under the same preprocessing conditions as in 3 prior studies: Dataset 1 (DS1) from [5], with $N = 1100$ subjects, preprocessed using the Connectome Computation System pipeline, band-pass filtering and no global signal regression, and parcellated with the CC200 atlas; Dataset 2 (DS2) from [9], with $N = 1035$ subjects, preprocessed using the Configurable Pipeline for the Analysis of Connectomes, band-pass filtering and global signal regression, and parcellated with the CC200 atlas; and Dataset 3 (DS3) from [1], with $N = 870$ subjects, preprocessed using the same pipeline as in [9] but parcellated with the HO atlas.

The time-series for each ROI of each subject was standardized by subtracting the mean and dividng by the standard deviation and resampled to 2s intervals between time points to harmonize the sampling across acquisition sites. We augmented the dataset by a factor of 10 during training by extracting 10 randomly cropped windows with length $T = 90$ timepoints from each subject during each epoch. At test time, every possible window of 90 timepoints is extracted from the time-series data for each subject and input to the trained network. The predicted probability of ASD for a given subject was then computed as the proportion of windowed samples classified as ASD.

Demographic information included gender, age, handedness, full IQ, verbal IQ, performance IQ, and eye status during scanning. Missing IQ data were imputed based on other available IQ scores for the subject, where we approximated full IQ as the average of verbal IQ and performance IQ, and subjects with no available IQ scores were assigned scores of 100, which is the mean population IQ. Each demographic variable was standardized to lie in the range of [-1,1].

### 3.2 Experimental Methods

Models for classification of ASD vs. HC were trained for each subset of the ABIDE dataset. We compared and implemented the following models which have the same underlying LSTM baseline architecture and incorporate demographic information: the proposed demographic-guided attention network (DGA); the DGA network without the residual connection, i.e. using the computed context alone (DGA-C); the baseline LSTM network combined with separately processed demographic information through late fusion as proposed in [7] (DFuse);

the baseline LSTM network with the hidden state and cell state of the LSTM initialized based on the demographic information as proposed in [6] (DInit). Models were implemented in Keras, with 32 nodes for the LSTM. For regularziation, models were trained using a dropout layer before each fully connected layer (with 0.5 probability of node dropout). Optimization was performed using the Adam optimizer, with binary cross-entropy loss or with QDL as in Eq. 3 for DGA2, a batch size of 32, and early stopping based on validation loss and a patience of 5 epochs. DGA-based models were tested with 1 (DGA1) or 2 (DGA2) attention heads and QDL with $\lambda = 0.5$ (DGA2-QDL). In addition, we compared the original study for each dataset that used only imaging information.

To assess the implemented models, we used LOSO CV, repeating the CV 5 times and averaging the performance measures for each site across CV runs both with and without weighting by the number of subjects per test site. We chose the LOSO framework to better estimate the model generalizeability compared to the commonly employed stratified k-fold cross-validation, which gives overoptimistic results. We measured classification performance by computing the accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and area under the receiver operating characteristic curve (AUC). We tested for differences against the baseline LSTM model by comparing the performance for the same left-out sites using two-tailed paired t-tests with a significance level of 0.05.

We also evaluated functional networks that were attended to based on individual demographic factors by applying the Neurosynth decoder [15], which correlates over 14000 fMRI studies with 1300 descriptors. For the 2-head attention model with QDL loss, we computed the correlation between the demographic variable $d(i)$ and the context for functional network $f$ from each head $c_1(f)$ and $c_2(f)$ across the test ASD subjects. We analyzed the US and Yale site as their test accuracy was high ($> 75\%$) and they contained significant heterogeneity for the investigated demographic variables of age, gender, handedness, and full IQ. We then found the functional network $f$ that resulted in the largest difference in correlation values for the 2 heads. The binary mask of the functional network of interest was then input to Neurosynth to assess neurocognitive processes associated with different modes of heterogeneity in ASD.

### 3.3 Classification Results

Classification results for each dataset are summarized in Tables 1-3. The results using the method from the original study for DS1 and published in the original study for DS2 and DS3 use only fMRI data and are shown in the first entry. We notice that generally, the fusion model DFuse and LSTM initialization model DInit do not perform significantly differently from the baseline LSTM model, particularly for DS3. The DGA-based models that use the context alone as the input to the FC (DGA1-C and DGA2-C) tend to perform about the same (DS1) or better (DS2 and DS3) than the non-DGA models. Adding in the residual connection for DGA1 and DGA2 results in similar (DS1 and DS2) or better (DS3) results than the DGA-C models. Finally, the DGA2-QDL model resulted in the top performance for DS1 and DS2 as measured by accuracy and AUC.

Table 1: DS1 Classification Results (N = 1100, 48.1% ASD)

| Model | Leave-One-Site-Out | | | | Weighted by # Subjects/Site | | |
|---|---|---|---|---|---|---|---|
| | Mean (Std) ACC (%) | Mean (Std) TPR (%) | Mean (Std) TNR (%) | Mean (Std) AUC | Mean (Std) ACC (%) | Mean (Std) TPR (%) | Mean (Std) TNR (%) |
| Orig (LSTM) [5] | 63.4 (0.7) | 60.9 (1.2) | 66.2 (0.5) | 0.695 (0.006) | 65.0 (0.7) | 61.3 (1.3) | 68.4 (1.2) |
| DFuse [7] | 63.3 (1.2) | 55.7 (3.3) ◦ | 70.7 (2.5) * | 0.701 (0.017) | 65.4 (1.3) | 57.7 (3.3) | 72.5 (3.3) |
| DInit [6] | 65.4 (0.6) * | 60.7 (1.2) | 69.9 (0.6) * | 0.709 (0.006) | **67.1 (0.7)** * | 62.6 (2.4) | 71.3 (2.5) * |
| DGA1-C | 64.4 (0.7) | **62.5 (0.6)** | 66.3 (1.5) | 0.710 (0.009) | 65.9 (0.5) | **63.5 (1.4)** | 68.1 (0.9) |
| DGA2-C | 64.3 (1.2) | 56.2 (2.3) ◦ | 71.8 (3.0) * | 0.703 (0.006) | 65.8 (1.1) | 57.3 (1.7) ◦ | 73.8 (3.0) * |
| DGA1 | 63.8 (0.9) | 61.5 (2.9) | 66.1 (1.9) | 0.702 (0.009) | 65.7 (1.1) | **63.5 (3.0)** | 67.7 (2.7) |
| DGA2 | 64.8 (2.4) | 56.1 (3.8) | **73.1 (2.1) *** | 0.710 (0.011) | 66.3 (1.6) | 57.1 (3.2) ◦ | **74.8 (2.5)** * |
| DGA2-QDL | **65.5 (0.8)** * | 59.1 (2.3) | 72.0 (2.4) * | **0.711 (0.006)** | 66.8 (0.7) * | 60.7 (1.3) | 72.4 (1.9) |

* Significantly different compared to LSTM with no demographic input ($p < 0.05$), with larger mean value.

◦ Significantly different compared to LSTM with no demographic input ($p < 0.05$), with smaller mean value.

Table 2: DS2 Classification Results (N = 1035, 48.8% ASD)

| Model | Leave-One-Site-Out | | | | Weighted by # Subjects/Site | | |
|---|---|---|---|---|---|---|---|
| | Mean (Std) ACC (%) | Mean (Std) TPR (%) | Mean (Std) TNR (%) | Mean (Std) AUC | Mean (Std) ACC (%) | Mean (Std) TPR (%) | Mean (Std) TNR (%) |
| Orig† [9] | 65 (1.5) | 69 (2.6) | 62 (2.7) | - | 65.4 (1.3) | 68.1 (2.6) | 62.3 (2.6) |
| LSTM [5] | 63.6 (0.5) | 55.2 (1.6) | 71.9 (0.6) | 0.709 (0.006) | 65.6 (0.6) | 58.2 (1.7) | 72.7 (0.9) |
| DFuse [7] | 65.5 (0.9) * | 57.1 (0.6) | 73.5 (1.6) | 0.713 (0.006) | 67.2 (0.6) | 61.2 (1.2) | 72.8 (1.0) |
| DInit [6] | 65.8 (0.8) * | 58.1 (0.4) | 72.9 (1.4) | 0.720 (0.009) | **67.5 (1.1)** * | 61.8 (1.6) * | 72.9 (3.2) |
| DGA1-C | 65.6 (1.7) * | 61.1 (1.6) | 69.6 (1.1) | 0.713 (0.011) | 66.8 (1.6) | **64.1 (2.0)** * | 69.3 (1.9) |
| DGA2-C | 65.8 (0.9) * | 52.6 (2.4) | **78.3 (1.7)** * | 0.719 (0.009) | 67.2 (1.2) * | 55.9 (2.4) | **78.0 (0.8)** * |
| DGA1 | 66.1 (1.5) * | **61.3 (2.5)** * | 70.4 (1.4) | 0.719 (0.011) | 67.4 (1.7) * | 63.6 (2.3) * | 70.9 (1.7) |
| DGA2 | 65.5 (1.0) * | 54.3 (1.5) | 76.5 (1.4) * | 0.716 (0.015) | 67.1 (1.4) | 57.6 (1.3) | 76.1 (2.3) * |
| DGA2-QDL | **66.4 (0.4)** * | 58.0 (1.9) * | 74.2 (2.0) | **0.722** (0.006) | 67.4 (0.5) * | 61.3 (1.7) * | 73.1 (1.9) |

† Values taken from the literature, reflecting one round of LOSO CV.

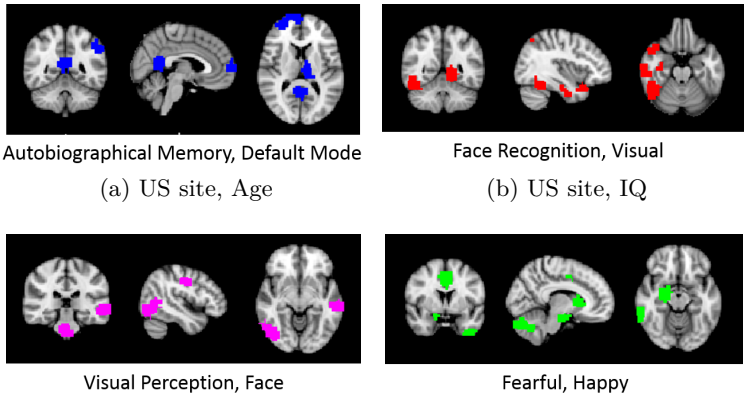
* Significantly different compared to LSTM with no demographic input ($p < 0.05$).

To better understand the performance over all the datasets, we scored each model by the number of performance measures that significantly improved over the baseline LSTM, minus the number of measures that significantly worsened compared to baseline, plus the number of top ranked measures. The models ranked in order of increasing performance was then DFuse, DGA2-C, DInit, DGA1-C, DGA1, DGA2-QDL, DGA2. Thus, DGA-based models generally performed better than other demographic models; 2-headed attention was generally better than 1; and the proposed residual connection for using the context as a bias to the LSTM outputs generally performed better than using the context alone. The reason for DGA2-QDL's lower ranking is due to the performance on DS3; we posit that the lower number of subjects in this dataset led to less heterogeneity, thus making it difficult to find two disparate attention modes, which QDL is trying to recover by minimizing the projection space similarity.

Table 3: DS3 Classification Results (N = 860, 46.1% ASD)

| Model | Leave-One-Site-Out | | | | Weighted by # Subjects/Site | | |
|---|---|---|---|---|---|---|---|
| | Mean (Std) ACC (%) | Mean (Std) TPR (%) | Mean (Std) TNR (%) | Mean (Std) AUC | Mean (Std) ACC (%) | Mean (Std) TPR (%) | Mean (Std) TNR (%) |
| Orig† [1] | 63.6 (6.2) | 59.8 (10.3) | 66.7 (12.8) | - | - | - | - |
| LSTM [5] | 63.8 (0.4) | 50.3 (1.9) | 75.5 (1.4) | 0.694 (0.012) | 65.3 (0.6) | 53.9 (2.3) | 75.0 (1.9) |
| DFuse [7] | 65.6 (1.6) | 52.7 (1.8) | 76.4 (3.0) | 0.714 (0.007) | 67.1 (0.8) * | 56.9 (2.5) | 75.8 (1.4) |
| DInit [6] | 64.5 (1.2) | 50.5 (2.2) | 76.4 (1.9) | 0.702 (0.013) | 66.3 (0.8) | 55.5 (2.4) | 75.5 (2.4) |
| DGA1-C | 65.5 (1.2) * | **55.3 (1.6)** * | 74.5 (2.1) | 0.708 (0.010) | 66.8 (0.9) * | 59.0 (2.4) * | 73.5 (3.4) |
| DGA2-C | 65.9 (1.6) | 52.6 (1.7) | 78.6 (2.0) | **0.717 (0.014)** | 67.2 (1.2) | 55.2 (1.8) | 78.6 (1.5) |
| DGA1 | 65.8 (0.1) * | **55.3 (1.1)** * | 75.2 (1.2) | 0.712 (0.006) | 66.8 (0.7) * | **59.1 (1.8)** * | 73.3 (1.5) |
| DGA2 | **66.8 (1.0)** * | 51.2 (2.1) | **80.0 (2.7)** * | 0.714 (0.005) | **68.0 (1.0)** * | 54.0 (2.5) | **80.0 (2.7)** * |
| DGA2-QDL | 66.0 (1.1) | 53.5 (1.4) | 76.8 (3) | 0.709 (0.006) | 67.0 (0.9) * | 57.6 (2.9) | 75.2 (3.3) |

† Values obtained from corresponding author of [1], reflecting one round of LOSO CV.
* Significantly different compared to LSTM with no demographic input ($p < 0.05$).




(a) US site, Age (b) US site, IQ

(c) Yale site, Gender (d) Yale site, Handedness

Fig. 2: Functional networks from the DGA2-QDL model trained on DS2 which had largest difference between the correlations of the listed demographic variable with the two attention measures $c_1(f)$ and $c_2(f)$ for ASD subjects in the listed test site. The top associated cognitive functions decoded by Neurosynth for each network are shown.

### 3.4 Demographic-guided Heterogeneity of Functional Processing

We explored the functional networks from the best model for DS2, DGA2-QDL, that corresponded to the most diverse outputs by the 2 attention heads. These different modes of the model's response to a functional network may correspond to potentially different mechanisms of ASD pathophysiology. Resulting functional networks and the top 2 associated Neurosynth cognitive terms are shown in Fig. 2. The functional networks highlight regions that are often associated with ASD (e.g., Fig. 2(b) and (c), visual perception and face processing [10]), and are also potentially associated with the demographic variable of interest (e.g., Fig. 2(a), default mode network changes with age [8]).

## 4 Conclusions

We have presented a novel demographic-guided attention mechanism for modeling the heterogeneity in neuropathophysiology of ASD. We achieved higher ASD classification performance on several ABIDE datasets and preprocessing conditions under a leave-one-site-out cross-validation framework, demonstrating improved generalization to data from new imaging sites. The success of having multiple attention modes for modeling the different neural mechanisms associated with ASD may help partially explain some of the conflicting results in the ASD literature (e.g., hyper- vs. hypo-connectivity), as our classification models improve once we account for the heterogeneity of the disorder.

## References


1. Abraham, A., Milham, M.P., Martino, A.D., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G.: Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. Neuroimage **147**, 736–745 (2017)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR 2015 (2015)
3. Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., ..., Bellec, P.: The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. In: Neuroinformatics (2013)
4. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., ..., Milham, M.P.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry (2014)
5. Dvornek, N.C., Ventola, P., Pelphrey, K.A., Duncan, J.S.: Identifying autism from resting-state fmri using long short-term memory networks. In: MLMI 2017. LNCS 10541 (2017)
6. Dvornek, N.C., Yang, D., Ventola, P., Duncan, J.S.: Learning generalizable recurrent neural networks from small task-fmri datasets. MICCAI 2018 (LNCS 11072) (2018)
7. Dvornek, N.C., Ventola, P., Duncan, J.S.: Combining phenotypic and resting-state fmri data for autism classification with recurrent neural networks. In: ISBI (2018)
8. Fair, D.A., Cohen, A.L., Dosenbach, N.U.F., Church, J.A., Miezin, F.M., Barch, D.M., Raichle, M.E., Petersen, S.E., Schlaggar, B.L.: The maturing architecture of the brain's default network. Proceedings of the National Academy of Sciences **105**(10), 4028–4032 (2008)
9. Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., Meneguzzi, F.: Identification of autism spectrum disorder using deep learning and the abide dataset. Neuroimage Clin. (2018)
10. Kaiser, M., Hudac, C., Shultz, S., Lee, S., Cheung, C., Berken, A., ..., Pelphrey, K.: Neural signatures of autism. Proc Natl Acad Sci U S A (2010)
11. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence (2016)
12. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: The 28th International Conference on Machine Learning (2011)
13. Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Moreno, R.G., Glocker, B., Rueckert, D.: Spectral graph convolutions for population-based disease prediction. In: MICCAI 2017. pp. 177–185. LNCS 10435, Part III (2017)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: 31st Conference on Neural Information Processing Systems (NIPS 2017) (2017)
15. Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wagerss, T.D.: Large-scale automated synthesis of human functional neuroimaging data. Nature Methods (2011), www.neurosynth.org