**Multiple Deep Learning Architectures Achieve Superior Performance Diagnosing Autism Spectrum Disorder Using Features Previously extracted from Structural and Functional MRI**

*Authors*

**Abstract:**

**Introduction:**

The diagnosis of Autism Spectrum Disorder (ASD) is dependent upon a subjective, time-consuming evaluation of behavioral tests by an expert clinician specializing in neurodevelopmental disorders. Since such expertise is not available at many clinics and the diagnosis can differ depending on the evaluator, there is growing interest in automating accurate diagnoses to increase its availability and reproducibility while reducing subjectivity. Non-invasive functional MRI (fMRI) and structural MRI (sMRI) characterize brain connectivity and shape and may be used to inform diagnoses and democratize medicine. Existing diagnosis automation studies that employ neuroimaging data are limited in two ways. First, they typically focus on one proposed predictive model and do not optimize comparable methods to the same degree as the proposed method. Second, they often require access to raw image data. Sharing raw patient images is problematic due to concerns for patient identifiability, while derived image features such as volumetry and functional connectivity are readily sharable. Elucidating the comparative performance of multiple deep learning methods and shallow machine learning (ML) methods would be of great significance to guide the image analysis community. Furthermore, successful construction of predictive models, such as deep learning models, from fMRI requires addressing key choices about the model's architecture, including the number of layers and number of neurons per layer. Meanwhile, deriving functional connectivity (FC) features from fMRI requires choosing an atlas with an appropriate level of granularity. Therefore, this research systematically compares the performance of classifiers to diagnose a subject as autistic or healthy. In total 12 classifiers are compared from 3 categories including: 6 non-linear shallow ML models, 3 linear shallow models, 3 deep learning models. Each classifier is similarly trained, optimized with random search-based optimization, and evaluated using over 900 subjects from the IMPAC database [1] using the database's derived anatomical and functional features. When evaluated with an area under the receiver operating characteristic curve (AUROC) performance metric, results include: (1) amongst the shallow learning methods, linear models outperformed nonlinear models, agreeing with [2]. (2) Deep learning models outperformed shallow ML models. (3) The best model was a dense feedforward network, achieving 0.80 AUC which compares to the 0.79+/-0.01 AUC average of the top 10 methods from the IMPAC challenge [3]. (4) When tested on two external datasets that were not used in training, the models achieved an AUC of up to 82%.

Once an accurate diagnostic model has been built, it is vital to determine which features are predictive of ASD and if similar features are learned across atlas granularity levels. Identifying new important features extends our understanding of the biological underpinnings of ASD, while identifying features that corroborate past findings and extend across atlas levels instills model confidence. To identify aptly suited architectural configurations, probability distributions of the configurations of high versus low performing models are compared. (5) Additional results show the highest performing models use between 2-4 hidden layers and 16-64 neurons per layer, granularity dependent. To determine the effect of atlas granularity, connectivity features are derived from atlases with 3 levels of granularity and important features are ranked with permutation feature importance. (6) Connectivity features identified as important across all 3 atlas granularity levels include FC to the supplementary motor gyrus and language association cortex, regions whose abnormal development are associated with deficits in social and sensory processing common in ASD. Importantly, the cerebellum, often not included in functional analyses, is also identified as a region whose abnormal connectivity is highly predictive of ASD. These results demonstrate that even when using features derived from imaging data, deep learning methods can provide additional predictive accuracy over classical methods. These results also indicate important regions to include in future studies of ASD, help assist in the selection of network architectures, and help identify appropriate levels of granularity to facilitate the development of accurate diagnostic models of ASD.

***Background:***

Autism spectrum disorder (ASD) is a common psychiatric disorder characterized by social and communication deficits and a restricted pattern of interests [4]. It is known that individuals with ASD have altered neuroanatomy and connectivity, though the full extent of these relationships has not been fully elucidated. Currently, the diagnosis of ASD is a subjective process that requires an expert in neurodevelopmental disorders who may be unavailable at many clinics. Non-invasive imaging captures structural and functional aspects of brain development that are promising for an automated machine learning (ML) based diagnosis. Such automated approaches would reduce subjectivity and increase reproducibility and availability of the diagnosis. Existing literature on automated diagnostics are limited in two ways. First, in these studies, just one category of predictive model is typically proposed and fully optimized; making comparisons to comparable methods biased. Second, they often depend on access to raw image data. Sharing patient images can be problematic due to concerns for patient identifiability. However preprocessed data, such as volumetry and functional connectivity are more easily shared. The Paris IMPAC Autism Challenge [1] is one such sharable dataset containing the derived features from structural MRI (sMRI) and resting state functional MRI (rs-fMRI).

***Previous Works***

Autism Spectrum Disorder (ASD) is among the most common developmental disorders affecting 1 in 160 children annually and is characterized by abnormal neurological development [5]. Accurate Expert ASD diagnosis is less available than desired. Consequently, there is growing interest in the development of an accurate, objective, fast, and reproducible diagnostic approach. One such approach uses machine learning combined with one or both of functional MRI (fMRI) and structural MRI (sMRI) which can measure anatomical and functional alterations manifest in ASD [6, 7]. This approach is particularly promising when the imaging is used as the input to train a machine learning model to predict whether the subject has ASD or is a typically developing subject (e.g. a healthy control).

Prior work has been performed on both the IMPAC [\*\*\*cite\*\*\*] and ABIDE [\*\*\*] datasets. For multi-site accuracy and accuracy on a held-out test site, accuracies of leading works range from mid 60% to 80%, depending on the features used and the means of measuring accuracy. The results of the IMPAC challenge have the top 10 models (the majority being ensembled linear models) achieving 0.79+/-0.01 AUROC [\*\*\*cite\*\*\*]. There are few notable uses of the entire ABIDE dataset with a fully held-out test set. Sen uses both sMRI and fMRI to construct an SVM model and achieves a hold-out test set accuracy 64.3% on the ABIDE dataset, using all sites [10]. Ghiassian also uses both sMRI and fMRI to construct an SVM, achieving a hold-out accuracy of 65% on ABIDE, all sites [9]. Studies that report cross-validation accuracy include those by Khosla, who uses rsfMRI in a CNN model. They achieve a cross validation accuracy on ABIDE1 of 72%, and 72% on ABIDE2 using all sites [7]. Huang, using rsFMRI features and an SVM, achieves a leave-one-out cross-validation accuracy of 79.35% on all of ABIDE [15]. Parisot uses rsFMRI and a Graph-CNN to achieve a cross-validation accuracy of 70.4% on all of ABIDE [6]. Accuracies of models trained on a few imaging sites or a single site have been higher. Li used fMRI and a deep-learning SVM hybrid to achieve a 70% cross-validation AUROC on ABIDE using several sites [11]. Wang used fMRI and a sparse-MVTC to achieve a nested cross validation accuracy of 72.6% on the NYU site and a nested cross validation accuaracy of 71.4% on the UM site of ABIDE [12]. Kam used rsfMRI and Boltzmann machines to achieve a cross-validation accuracy of 80.82% on ABIDE, using the UM site only [13]. Kazeminejad used rsFRMI, graph theory and SVM on subgroups clustered by age. They achieved a cross-validation accuracy of up to 95% on one subgroup using all sites, though most subgroup cross-validation accuracies fell in the 70-80% range [14]. Though most of the above studies use the large, public ABIDE dataset, similar results have been achieved with others. For example, Yamagata used rsfMRI and logistic regression to achieve a leave-one out cross-validation AUROC of 78% AUROC on in-site dataset [16], and Dekhil used rsfMRI and an SVM method to achieve a cross-validation AUROC of 92.18% on the NDAR dataset [17].

The gaps in the research above are several fold: First, there is limited formal comparison across multiple model types and imaging resolutions. Second, the results are not tested on an external dataset, the gold standard of comparison. Finally, there is little work put into what imaging features are learned across multiple models and how robust these features are across scales. The limited research aimed at understanding how deep learning methods compare in performance to shallow ML methods on such datasets with pre-derived features is addressed here, along with testing on an external dataset and a thorough evaluation of the features learned across models and scales.

***Contributions***

We elucidate the comparative performance of model categories across a large sharable dataset and perform a systematic analysis across a breadth of ML models and parcellations in an unbiased fashion. In order to perform a fair comparison, in this study each model is similarly hyperparameter optimized using a random search-based approach. Identical randomly chosen cross-validation splits are used to train each model, ensuring similar training opportunities for each model. This is done to prevent bias in the model selection procedure where the machine learning engineer carefully optimizes only a subset of a small number of tested models. Finally, we thoroughly evaluate the effectiveness of our hyperparameter optimization and analyse what features were learned across models and across resolutions of parcellations.

The primary contributions of this work are two-fold:

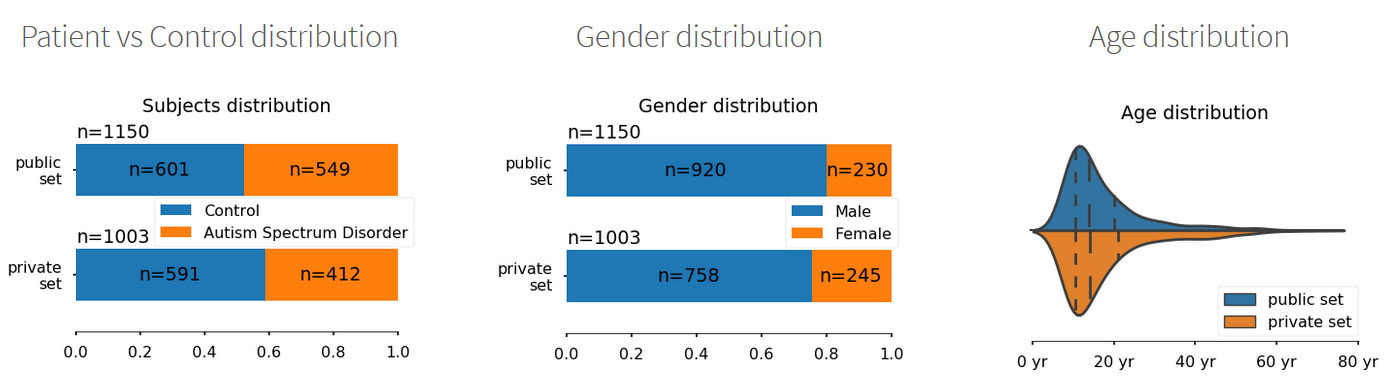
*First,* the study provides a systematic comparison of 3 broad categories of methods: linear and non-linear shallow ML models and deep learning models and assesses their relative performance. The study examines the relative performance of anatomical features, functional features and their combination and provides evidence of their level of complementarity. Evidence for the effective level of granularity for deriving regional features from whole brain parcellations is obtained by comparing 7 atlases. Therelative performance of 12 individual classifiers is compared and recommendations for ASD diagnosis is made for a specific winning deep learning model, which achieves greater performance than all other tested models.

*Second*, this work extracts information from the deep learning hyperparameter optimization, allowing us to identify configurations that lead to high performance and whether search ranges were adequate to isolate local performance maxima. This work also determines the most important FC features used by each model through permutation feature importance and compares these to regions known to be affected by ASD, which would help grow confidence that the models have learned appropriately. Finally, this work compares the discovered features across three levels of brain-atlas granularity. Features learned in common across models trained from different granularities can further corroborate their importance, and potentially identify novel features warranting further investigation.

**Materials and Methods:**

***Materials***

This study uses 915 subjects from the IMPAC dataset [1] that received both sMRI and rs-fMRI and were identified by the IMPAC organizers as having satisfactory images. The focus of this study is the comparison of two-category classifiers for diagnosing ASD or healthy control. The IMPAC dataset includes the clinical diagnosis (the classifier target) for which there were 418 ASD patients and 497 subjects designated as a healthy control. Structural MRI (sMRI) and resting state functional MRI (rs-fMRI) was acquired for each subject. Figure 1 illustrates how the features were derived from the MRI. From the sMRI, 207 features were extracted, including volumes of cortical and subcortical structures, cortical thickness, and area per region of interest (ROI) defined by the Desikan-Killiany gyral atlas [18]. From the rs-fMRI, functional connectivity matrices were derived. For this derivation, the rs-fMRI was first parcellated into ROIs using seven different atlases including: atlases (1-3) The BASC Atlas, whose regions are defined by k-means clustering of stable coherent groups [19] for k=64, 122, and 197 ROIs, atlas (4) the Craddock atlas, which defines 249 ROIs by coherence of local graph connectivity [20], atlas (5) the Harvard-Oxford Anatomical atlas, which defines 69 ROIs using anatomical features, atlas (6) the MSDL atlas, which has 39 ROIs defined by correlations of spontaneous activity [21], and atlas (7) the Power atlas [22], which is defined by local graph-connectivity into 264 ROIs. The rs-fMR time signals from each region were converted into a connectivity matrix by projection into tangent space, a procedure which captures subject-specific variations from one or more groups [23]. Clinical data including patient gender and age were also collected. Demographics are shown in Table \*\*\* below.

[ \*\*\*temporary image from the challenge website]

*Table \*\*\* Demographics of IMPAC subjects.*

For independent model confirmation, we also used sMRI and fMRI data from the ABIDE study [cite]. 176 subjects from the NYU site of part I of the ABIDE study and 105 subjects from the NYU site of part II of the ABIDE study were processed with an in-house processing pipeline with advanced motion correction [cite motion correction paper]. sMRI features were derived as described above for the IMPAC study using the freesurfer package [cite]. The sMRI extracted features were normalized by site. The tangent-space embedding model was fit on the entire IMPAC dataset, and the TSE values per-ROI timeseries for the ABIDE dataset were calculated. Due to the non-gaussian distribution of TSE values, the TSE values calculated on the ABIDE dataset were inverse-robust scaled [cite method?] to the same range as the IMPAC TSE values. Patient gender and age were also collected. Demographics of the NYU ABIDE subjects are presented in Table \*\*\* below.

[ \*\*\*insert demographic table here\*\*\*]

*Table \*\*\* Demographics of ABIDE I and ABIDE II NYU subjects.*

A picture containing clock

Description automatically generated

*Figure 1: Combinations of derived features used by the predictive models tested in this study. (A) The rs-fMRI was transformed into a symmetric connectivity matrix for each atlas. (B) Upper triangular elements of matrix were flattened into a 1D vector. (C) The sMRI was transformed into a vector of cortical/subcortical ROI volumes and cortical thickness features. In (D) the connectivity matrix vector is used as the sole input for the predictive model, in (E) both anatomical and connectivity derived feature vectors are appended and used, while in (F) the anatomical features are used as the sole input for the predictive model.*

***Data Partitioning***

Subjects were randomly partitioned with 80% assigned to a training set and 20% to a test set with the split having matching proportions of diagnosis (ASD/healthy) and gender (male/female). The test subjects were set aside and not used during training or model selection. The training set was further split into validation and training folds using a 3-fold stratified cross validation approach. To ensure fair subsequent model comparison, the same splits were used for all tested machine learning models.

***Model Construction:***

Systematic testing of a broad array of 12 machine learning classifiers was conducted. This included 6 non-linear shallow machine learning methods: a naïve Bayes classifier, a support vector machine with a Gaussian kernel, a random forest classifier, an extremely randomized trees classifier, adaptive boosting, and gradient boosting with decision tree base models; 3 linear shallow models:

a support vector machine with a linear kernel, logistic regression with ridge regularization, logistic regression with lasso regularization; and 3 deep learning approaches: a fully connected dense feedforward (FeedFWD) network, and a long-short term memory (LSTM) based recurrent neural network classifier (RNN), and the BrainNetCNN [24].

Classical models were constructed using the scikit-learn and XGBoost pakages, while the deep learning models used the keras, tensorflow, and caffe packages. The LSTM classifier uses a dense neural network atop an LSTM for classification like [25] which can succeed even on non-sequential fixed vector data, as suggested by [26]. The BrainNetCNN classifier is a graph-convolutional network classifier [24]. The models were trained on an NVIDIA Tesla p100.

***Random Search***

To fairly evaluate the models and avoid biasing the results, 50 points in hyperparameter space were randomly chosen for each model and each of these 50 models were trained across 3 cross-validation folds from the 80\% training set. The highest performing models were selected by mean AUROC across the cross-validation folds. To illustrate, consider the examples of simple and complex dense FeedFWD networks that were tested are illustrated in Table 1, in the left and middle columns respectively. In detail, for the models tested these hyperparameter points were randomly chosen from the following dimensions and ranges: *Random Forest: number of estimators [50, 5000], max nodes [5, 50]; Extremely Random Trees: number of estimators [50, 5000], max nodes [5, 50]; Adaptive Boosting: number of estimators [20, 2000], learning rate [0.1, 0.9]; Gradient Boosting: number of estimators [5, 5000], max depth [1, 10], subsampling fraction per tree [0.2, 0.8], fraction of columns per tree [0.2, 1], learning rate [0.01, 1]; SVM with Gaussian Kernel: C [0.0001, 10000], maximum iterations: [10000, 100000], gamma [0.01, 100]; SVM with Linear Kernel: C [0.0001, 10000], maximum iterations: [10000, 100000]; logistic regression with lasso regularization: C [0.0001, 10000], maximum iterations [1000, 100000]; logistic regression with ridge regularization: C [0.0001, 10000], maximum iterations [1000, 100000]; dense FeedFWD network: number of hidden layers [1, 3], layer width [16, 256]; dropout fraction [0.1, 0.6], L2 regularization coefficient [0.0001, 0.02]; LSTM: number of hidden layers [1, 3], layer width [16, 256], dropout fraction [0.1, 0.6], L2 regularization coefficient [0.0001, 0.02]; BrainNetCNN: number of hidden layers [0, 2], layer width [16, 64], dropout fraction [0.1, 0.6], ReLU slope for x<0 [0.1, 0.5].* The deep learning models used the leaky ReLU activation function, early stopping, the Nesterov ADAM optimizer, a batch size of 128, and the binary cross-entropy loss function.

Each of our 12 models types was trained on 15 different feature sets, for a total of 180 model type x feature set combinations. The feature sets contain measures of anatomical volume and functional connectivity from the IMPAC dataset. These feature sets included: (1) 207 measures of regional volume and thickness, (2-8) functional connectivity measured between regions defined by one of the 7 atlases described in the materials section above, (9-15) the union of the anatomical with one of the functional feature sets. All featuresets also included gender and imaging site as confounders. The model with the highest average area under ROC curve over the cross validation folds was selected as the best model for each model type x feature set combination. This model was then trained on all training data and evaluated on the held-out test set not used in training.

***Model Analysis***

The importance of understanding what features a model has learned cannot be understated. Thorough analysis of the most successful models after the unbiased random search has several different, complimentary purposes. First, by looking for cross-model concordance and concordance with features identified in independent studies, we can be more confident in our model learning appropriate and informative features. Second, we can use features that feature prominently within and across models to generate hypotheses about the most informative features for this diagnostic task, and even generate hypotheses about the pathophysiological changes associated with ASD. Finally, by understanding the important features found in each model and reducing its ‘black box’ nature, we will be more able to effectively bring these tools to the clinic and direct future endeavors.

Our best models achieved 75.4-80.4% ASD vs HC diagnosis accuracy using the BASC atlas [19] as opposed to 4 other atlases tested at 3 levels of granularity. This atlas' coarsest resolution contains 64 ROIs (Figure \*\*\* A) its medium-grained granularity has 122 ROIS (Figure \*\*\* B), while its fine-grained granularity has 197 ROIs (Figure \*\*\* C). The top 5 *DenseFFwd* models from validation AUROC trained on each BASC atlas are the subject of further analysis and interrogation because they performed best compared to other ML algorithms and well compared to the leading published results.

A close up of a logo

Description automatically generated

*Figure??? Brain parcellations (BASC [19]BASC) with varying granularity. (A) Coarse-grained with 64 ROIs, (B) medium-grained with 122 ROIs, (C) Fine-grained with 197 ROIs.*

***Model Ensembles***

Furthermore, given our generated set of high performing models from our random search, we also generated ensembles of models from the sets of high performing models. We exhaustively searched every combination of 3 chosen from \*\*\* models, \*\*\* atlases, and ensembling methods including \*\*\* and \*\*\*. The highest average AUROC from 3-fold cross validation for each ensemble was computed, but no further predictive benefit was found. As there was not increased predictive power from such ensembling approaches, we chose to instead delve further into the individual models. A more detailed discussion of both this approach and the results therin is found in the supplementary materials (Fig S1 and S2).

***High performing architectural configurations***

Hyperparameter searches generate a wealth of information. To gain insights from this information, kernel density estimates were computed for the models with the top 20% of performance and for the models with the lowest 20% of performance. This allows identification of architectural configurations that tend to produce high performing models and configurations that tend to produce low performing models. In addition, this analysis can indicate whether hyperparameter search ranges were adequate.

***Computation of feature importance***

The importance of each feature for each of the analyzed deep learning models was computed using permutation feature importance (PFI) [27]. In this approach, for a given trained model, each feature is permuted individually. Its feature importance, , is calculated as the z-score normalized mean decrease in AUROC: , between the performance before feature permutation () minus the performance after feature permutation (). PFI was chosen because it can be applied uniformly to different model and feature types.

To aid in the comparison of IMPAC connectivity features to the literature which often reports results in Brodmann areas (BA), the centroid of each ROI for each atlas was calculated and matched to the corresponding BA for cross-study comparison. The ROI-ROI connection can then be re-written as the closest BA-BA connection.

**Results**

*A screenshot of a cell phone

Description automatically generated*

*Figure 2: Performance of classifiers predicting the diagnosis of ASD versus healthy control. Performance is measured as the area under the ROC curve on held-out test data. Cooler colors indicate superior performance.*

The results of the random search are summarized in Figure \*\*\*, which shows the area under the ROC curve of different machine learning models predicting ASD vs healthy control on the test data that was held out from training. The models displayed had the highest average AUROC across the 3 cross-validation folds in training.

***The importance of feature set combination***

Comparing the 15 feature sets (rows of Figure 2), it can be observed that the anatomical features yielded the lowest prediction accuracy by area under the ROC curve, while the rs-fMRI functional connectivity features alone yielded models with higher predictive power than anatomical features. For functional connectivity data alone, the BASC atlas with any number of parcellations and the Power atlas generated models with more predictive power than other atlases. However, the combination of anatomical and functional features yielded models with even higher predictive power, suggesting their complementarity. The best performing models used the anatomical features with connectivity features from the Power atlas, Craddock atlas, or BASC atlas. Models trained on the Harvard-Oxford atlas connectivity data and volumetric data are notably lower performing, and models trained on the MSDL atlas were slightly better than those trained on the Harvard-Oxford atlas.

***The importance of model type***

The most accurate shallow machine learning algorithms were the SVM with either a Gaussian or linear kernel and the logistic regression with ridge regularization. The least successful methods were the random forest and extremely randomized trees. Both adaptive boosting and gradient boosting performed better than the random forest models, but overall did not perform as well as the linear methods. However, deep learning models performed even better than these shallow methods.

As shown in the columns towards the right of Figure \*\*\*, the deep learning methods performed higher than the other categories of models. The most successful deep learning algorithms were the dense FeedFWD network and LSTM, which outperformed other methods by a moderate margin. The BrainNetCNN model is only defined for functional connectivity input features, but on those features it performed lower than the other deep learning models with a performance similar to the linear models. Like the shallow methods, the deep learning methods performed best when using the combination of functional and anatomical features. The highest overall performance was the dense FeedFWD network, whose architecture is shown in Table 1, right column, using the rs-fMRI connectivity data with the BASC atlas with 122 ROIs and the sMRI volumetric data combined, achieving an AUC of 0.804. Other permutations using the BASC atlases, Craddock atlas, and Power atlas as training data for the dense FeedFWD network also performed well.

***Table 1:***

|  |  |  |
| --- | --- | --- |
| Simple Dense Network | Complex Dense Network | Highest Performing Dense Network |
| L2 Regularization: 2.3e-4  Dense 16 neurons  Dropout: 53% removed  Dense 16 neurons  Decision Layer 1 neuron | L2 Regularization: 2.3e-4  Dense 128 neurons  Dropout 18% removed  Dense 128 Neurons  Dropout 18% removed  Dense 64 neurons  Dropout 18% removed  Dense 42 neurons  Decision layer 1 neuron | L2 Regularization: 1.1e-4  Dense 64 neurons  Dropout 13% removed  Dense 64 neurons  Decision layer 1 neuron |

*Examples of dense FeedFWD network architectures tested in the random search. Hyperparameters shown include the regularization coefficient, \# of layers, \# of neurons/layer, and dropout fraction. Left column illustrates a simple network. Middle shows a complex network. Right column shows the architecture of the highest performing network.*

***Model Performances on Held-out Data***

Mensch et al. [28] reported high performance using deep learning networks for decoding brain activity to predict of the class of psychological stimuli presented in neuroimaging studies. This study focused on the classification of ASD versus healthy control and also demonstrated high performance using deep learning, adding to the evidence that deep learning is effective at classification from multidimensional neuroimaging data. The highest performing model in this study was a dense FeedFWD network which achieved 0.80 AUC, which is quite similar to the 0.79+/-0.01 AUC average of the top 10 methods recently reported from the IMPAC challenge [3]. Machine-learning classification of ASD from rs-fMRI and sMRI has been reported by Parisot et al. [6] and Khosla et al. [7]. These methods both employ novel convolutional neural networks to achieve state of the art performance of 70.4% and 73.3% accuracy respectively on the open source ABIDE dataset for ASD (\*\*\*cite ABIDE\*\*\*), but both depend on the raw imaging data.

***Hyperparameter Analysis***

***A close up of a piece of paper

Description automatically generated***

*Figure \*\*\* Kernel Density Estimates from the hyperparameter search reveals the density of top performing configurations (top 20%) shown in blue, and low performing configurations (bottom 20%) in orange. Densities of DenseFeedFwd configurations using the coarse BASC atlas (A), medium atlas (B), and fine atlas (C). Peaks of blue surfaces are marked with \*.*

Performance of a diagnostic predictive model can depend substantially on the choice of architectural configuration. In order to understand whether this is applies here, kernel density estimates were computed to estimate the probability distribution functions of the configurations of the top performing (top 20%) configurations and bottom performing (bottom 20%) configurations (Fig. \*\*\*).

Quantitatively, high performing *DenseFFwd* models tended to use 1-2 hidden layers with 64 neurons per layer versus 3-4 layers with 256 neurons for the low performing models *when using the coarse atlas* (Fig. \*\*\*A), 2 layers with 16-32 neurons vs. 3 layers with 128 neurons for the *medium-grained atlas* (Fig. \*\*\*B), and 3-4 layers with 16 neurons vs. 2 layers with 256 neurons for the *fine-grained atlas* (Fig. \*\*\*C). As the peaks of the high (blue) and low (orange) performing models are not proximal, this suggests that *configuration impacts performance substantially* (AUROC varied by 20% or more between high and low performing models). Also, we observe that the configurations of the top performing models, i.e. at the peaks in the blue surfaces, occur near the centers of the search ranges and not near the edges of the search space. This suggests that the search ranges have adequate coverage to discover good configurations.

***Model Performances on Unseen Dataset:***

To know if our machine learning models are truly generalizable, we test it on an external, similar autism dataset. Each of the top 5 Dense FeedFwd models using the combined structural and functional features derived with the BASC atlas with 64, 122, and 197 ROIs was tested on the NYU site of the ABIDE 1 and ABIDE 2 datasets. 10% of the testing data was used to tune the weights of the original model due to the known strong inter-site differences in the ABIDE dataset [\*\*\*cite\*\*\*]. Results of this external dataset validation is reported for the ABIDE I NYU site in *Tables \*\*\** andfor the ABIDE II NYU site in *Table \*\*\*.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Top PerformingModel on IMPAC | 2nd Best Model on IMPAC | 3rd Best Model on IMPAC | 4th Best Model on IMPAC | 5th Best Model on IMPAC |
| BASC atlas, 64 ROIs | 66.90+/-9.60 | 67.60+/-10.10 | 68.68+/-7.83 | 66.68+/-9.53 | 67.09+/-8.81 |
| BASC atlas, 122 ROIs | 70.88+/-9.27 | 70.82+/-9.24 | 69.57+/-9.51 | 71.35+/-9.20 | 68.15+/-8.38 |
| BASC atlas, 197 ROIs | 73.13+/-7.17 | 71.78+/-6.48 | 73.12+/-9.42 | 72.75+/-8.49 | 72.01+/-7.72 |

*Table \*\*\*. Model performances (AUROC) of the top performing IMPAC models on ABIDE I with 10% of data used for tuning.*

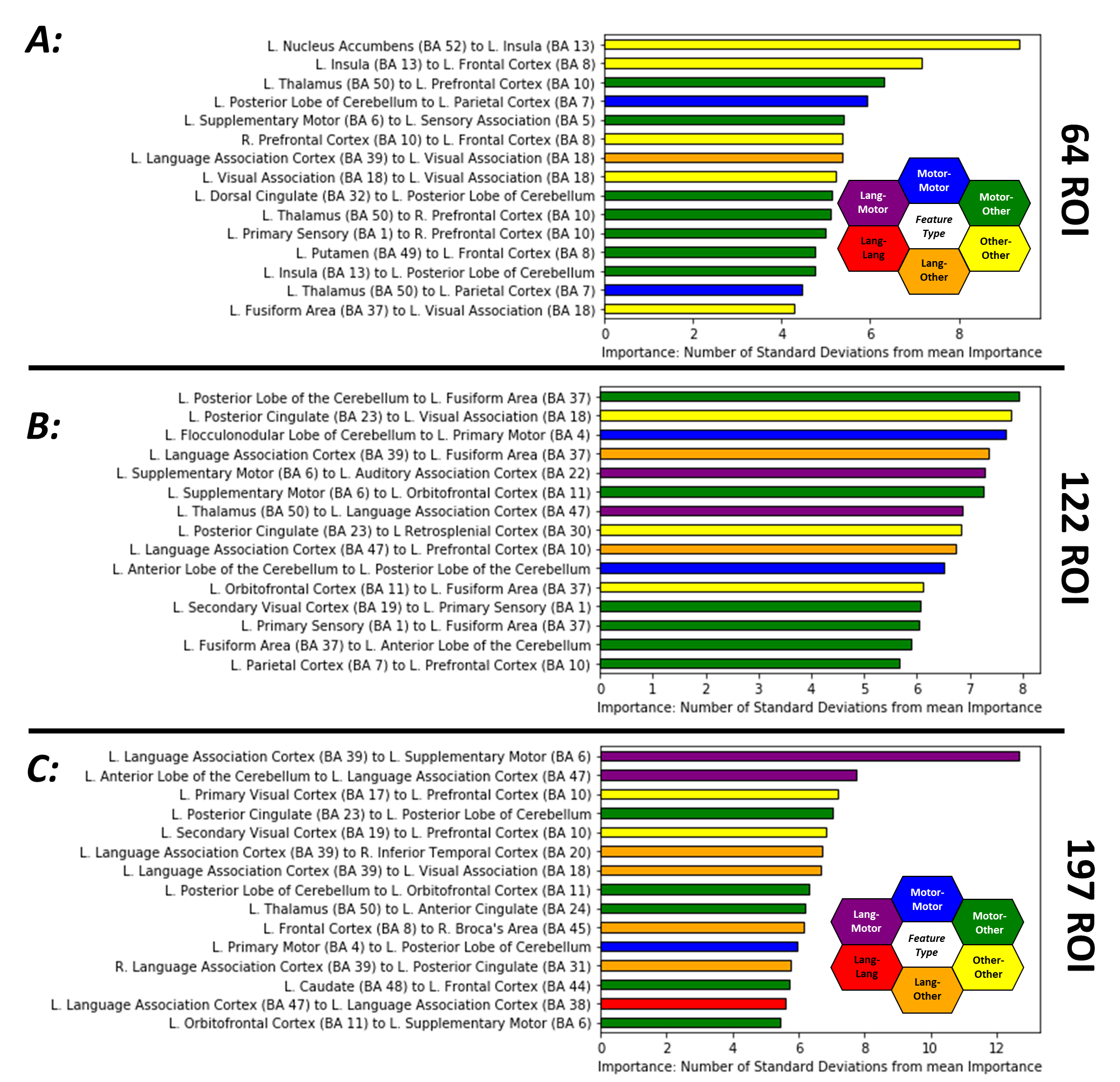
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Top Performing Model on IMPAC | 2nd Best Model on IMPAC | 3rd Best Model on IMPAC | 4th Best Model on IMPAC | 5th Best Model on IMPAC |
| BASC atlas, 64 ROIs | 78.57+/-9.54 | 81.10+/-11.52 | 77.94+/-10.18 | 78.88+/-11.85 | 76.67+/-10.52 |
| BASC atlas, 122 ROIs | 79.40+/-11.37 | 77.88+/-10.82 | 78.53+/-11.02 | 79.11+/-9.23 | 81.03+/-9.14 |
| BASC atlas, 197 ROIs | 81.94+/-10.52 | 79.35+/-11.00 | 80.87+/-7.95 | 79.78+/-11.5 | 80.53+/-9.78 |

*Table \*\*\*. Model performances (AUROC) of the top performing IMPAC models on ABIDE II with 10% of data used for tuning.*

In the ABIDE I dataset, 17-18 subjects were used to tune the model with the same early stopping criterion as used in the initial model training. In the ABIDE II dataset, 10-11 subjects were used to tune the model with the same early stopping criterion as used in the initial model training. The models were then tested on the other 90% of the dataset. This was repeated 10 times, with each block of 10% of the data being used to tune from the initial IMPAC weights and tested on the other 90 % once. The mean performance is reported in *Tables \*\*\* and \*\*\**. The models tested on ABIDE I achieved performances in the 70% range, while the models tested on the ABIDE II dataset achieved performances up to the low 80%’s.

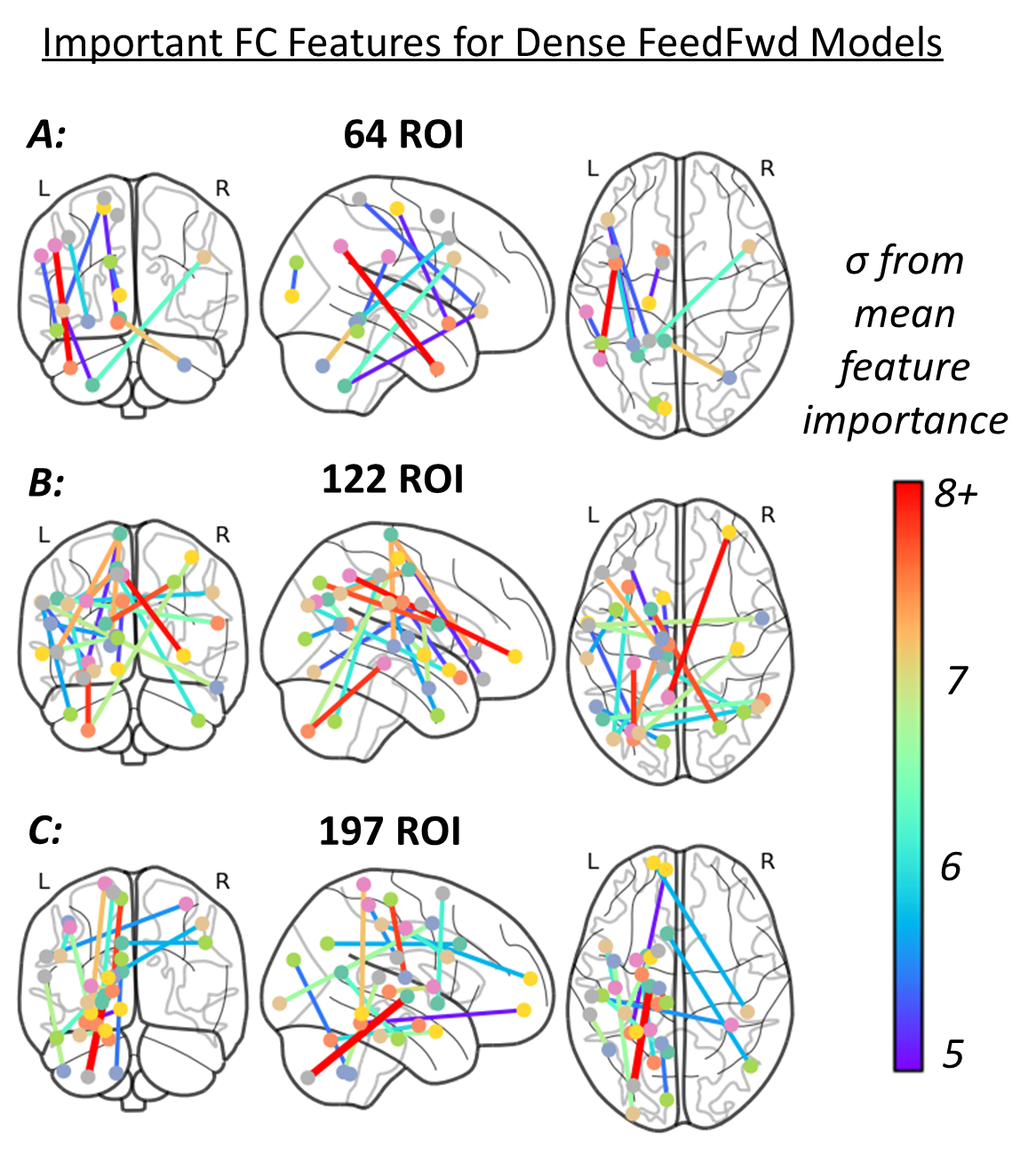
***Feature Importance***

The top 15 features ranked by their mean feature importance over the top 5 Dense FeedFwd models for each atlas granularity are shown in Fig. \*\*\*. The feature importance for the connectivity features are reported as the number of standard deviations from the mean calculated feature importance (z-score). The most important features for the ASD vs HC prediction for the model trained with 64 ROIs is shown in Fig. \*\*\*A, whereas Fig. \*\*\*B and Fig. \*\*\*C show the most important features for the models trained from 122 and 197 ROIs respectively. Color-coded functional labeling of features is shown to facilitate qualitative comparison. Motor, sensory, and language areas appear throughout the top features, while no structural features cortical thickness, volume, etc.) were among the top 15 most important discriminative features. Confounders were found to be significant features only in the 197 ROI models, and are not pictured. The important confounding variables included sex and 3 of the more than 30 sites.



*Figure ??? Important features for learned by highly accurate models for ASD diagnosis at each level of atlas granularity with coarse-grained atlas (A), medium-grained (B), and fine-grained (C). Each feature is the functional connectivity between two brain regions and is given a distinct color based on the function of these regions. Connections between sensorimotor ROIs are shown in blue, while connections between language ROIs are in red. Connections between regions that are neither motor nor language are in yellow. A connection between language (red) and motor (blue) ROIs is shown with an intermediate hue (i.e. purple) and similarly for other region function combinations.*

The three most important features for the *Dense FeedFwd* model trained from 64 ROIs were within the left association cortex and between association and somatomotor cortex. For the model trained using the medium-grained atlas (122 ROIs), the top 3 features were between association and somatomotor cortex, and within association and somatomotor cortex. For the fine-grained atlas (197 ROIs), the top three features were between somatomotor regions and language associated cortex and within association cortex.



*Figure ??? Neuroanatomical locations of the most important functional connectivity features and their relative importance. Features for the DenseFFwd model using the BASC atlas with coarse (A), medium (B), and fine (C) granularities. Features with z score >=5 are shown while their color indicates the number of standard deviations they are from the mean feature importance.*

The anatomical location of features with a z-score>=5 are shown in Fig. \*\*\*. The features are the same as those shown in Fig. \*\*\*. There is some overlap of feature importances across the different granularities tested.

**Discussion**

***Top Performing Models***

Within the model search itself, we can draw multiple conclustions: *First*, the results show that deep learning is still a valuable tool that can extract additional predictive power over shallow methods even when provided pre-extracted feature sets. *Second*, the subset of atlases that performed better is informative for ASD diagnosis. Across many different machine learning moralities, the functional BASC atlas, derived using a k-means clustering approach, performed very well. Its 122 and 197 ROI versions performed better than the 64 ROI version. This suggests the scale or granularity of neuroimaging-detectable changes in functional connectivity in ASD. Also, this suggests that k-means clustering and other graph-based clustering methods such as the Power and Craddock atlases, may be more suited to accurately elucidate functional connectivity changes in ASD than other parcellation methods. *Third*, the uniformly poor performance observed when models use purely anatomical features suggests that the deficits in ASD are reflected more by changes in functional connectivity than by changes in volume and cortical thickness. This finding is in agreement with results of previous studies [7](\*\*\*more\*\*\*)

However, the fact that in general, combining anatomical features with functional connectivity features tended to improve model performance across model categories, supports the notion that the information is complementary and should be combined to maximize predictive accuracy. Considering the classical shallow machine learning models, linear models were more successful at classifying autistic vs non autistic patients. When anatomical and connectivity data were combined, the Harvard-Oxford atlas generated the least predictive models, possibly because the ROIs in this atlas are based on anatomical features rather than on functional regions, hence the atlas is less complementary to the anatomical volume and thickness features. However, the fact that in general, combining such anatomical features with functional connectivity features tended to improve model performance across model categories, supports the notion that the information is complementary and should be combined to maximize predictive accuracy.

***Model Features learned***

Our hyperparameter search analysis revealed that the highest performing models used between 2 and 4 hidden layers with 16-64 neurons per layer, with the optimal number of layers increasing with granularity and the optimal number of neurons/layer decreasing with increasing granularity.

Several regions were consistently found to have altered FC across the 3 accurate models we examined. When considering the top 15 most important features for each model, *supplementary motor regions* (BA 6) were involved in 11 FC features across all three models, the *posterior cerebellum* is involved in 8 FC features, and *language association cortex (BA 39) and secondary visual cortex (BA 19)* are involved in 6 FC features across all three models. Similarly, the *left anterior cerebellum* is involved in 5 FC features, and Brodmann areas 54, 49, 48, 18, and 10 are all involved in 4 FC features across the three models. Additionally, we observe that many of the nodes implicated at a coarser resolution are also important at finer granularities (Fig. \*\*\*). For example, the FC from the supplementary motor area to the cerebellum and itself are important in the coarse atlas, while the supplementary motor FC to the putamen, cerebellum, hippocampus, and prefrontal cortex, and sensory cortex are important in the fine-grained atlas. Some of the differences in important edges across granularity may be explained by movement of the apparent regional centroid when a region is fine-grained. However, it also may be that these patterns of connectivity emerge only at specific scales. Further investigation into features variable across resolutions is warranted. That some features recur at multiple resolutions bolsters confidence in their importance and suggests that even higher granularity may be warranted to further elucidate biological underpinnings.

Many of the features identified here are in agreement with alterations reported previously including the significantly altered DMN connectivity, [29–31], connectivity in visual areas [31–34], motor and supplementary motor Connectivity [33], connectivity in somatosensory association areas [34, 35], and connectivity in the prefrontal cortex [29, 30, 35] in individuals with ASD.

Importantly, our analysis also indicates that the FC with the cerebellum, including both the anterior and posterior aspects, are important diagnostic predictors. Moreover, these cerebellar features are important across *all levels of granularity examined* (from the BASC atlas at 64, 122, and 197 ROIs). These consistent discriminatory connections lie between the cerebellum and motor areas as well as between the cerebellum and frontal cortex, regions that pertain to sensory processing and social behavior, well known to be altered in ASD. The altered FC between the cerebellum and frontal and sensorimotor cortices as a marker of ASD *has received little attention in the literature*, as the cerebellum is often not included in functional analyses. We suggest that these connections are areas worthy of further investigation and research.

***Limitations***

There are several limitations in this study. First, this dataset only provides us a binary diagnosis and, as ASD is well-known to be a spectrum disorder, a dataset with finer gradiations of diagnosis would be expected to provide a more precise diagnostic model. Second, other methods than PFI should be applied to the models, such as layer-wise relevance propagation, to further explore the learned abstractions.

**Conclusions**

***Takeaways***

In summary, this study provides insights into the comparative performance of three categories of widely used machine learning models, including both linear and non linear shallow models as well as deep learning models for the important task of automating diagnosis for Autism Spectrum Disorder. It provides insights into the combination of anatomical and functional features that are most useful for diagnosis of ASD and demonstrates that their combination is most appropriate. The study also demonstrates that a finer level of granularity in whole brain parcellation with roughly 120 ROIs outperforms coarser parcellations. Lastly the study shows that a dense FeedFWD network outperforms other models even when features are pre-extracted from MRI and attains highly accurate diagnosis compared to previously published methods. In the future we aim to continue to improve upon automated classification performance in ASD and other neuropathologies.

In conclusion, this work has characterized the architectural configurations that lead to high performing Deep Learning diagnostic models for ASD across 3 levels of granularity. This work has also identified the most important features using permutation feature importance analysis. The feature analysis identified new regions such as the anterior and posterior cerebellum with diagnostic importance and identified features in agreement with neuroanatomical regions previously implicated in ASD. That these regions were overlapping across 3 levels of regional granularity bolsters confidence that the models have discovered true discriminative features that may generalize well to new datasets and the clinic. We look forward to extending this work with further development to include additional regions such as the brainstem, further testing, and clinical translation.

**Supplementary**

***Concatenated models***

An additional experiment where two featuresets were combined and used as input into a Dense FeedFwd model was conducted. The models using the combined features performed as well as the top singular models, and due to the additional complexity at no benefit to performance, were left out of subsequent analyses.

Machine generated alternative text:
BASC Atlas 
with 64 
regions 
BASC Atlas 
with 122 
BASC Atlas 
with 197 
regions 
76.8&1. 
1 
73.6B. 
3 
75.5&2.9 
1 
78.8&0. 
5 
75.6B. 
5 
Craddock 
Atlas with 
249 regions 
74.1&2.6 
74.4&2.8 
73.1&2.9 
76.1&1.7 
76.4&1.2 
74.9&2.5 
Harvard- 
Oxford Atlas 
MSDL Atlas 
with 39 
regions 
74.5&0.8 
75&1.3 
76.6to.8 
74&1 
67.8&1.1 
75&2.8 
Power Atlas 
with 264 
regions 
75.2&2.2 
75.6&2.7 
74.7&2.7 
72.3&3.7 
72.4&2.6 
73.8&2 
9 
7 
7 
9 
9 
with 69 
regions 
74.2&0. 
8 
7 
regions 
74.4&2. 
7 
9 
2 
2 
BASC Atlas with 
64 regions 
BASC Atlas with 
122 regions 
BASC Atlas with 
197 regions 
Craddock Atlas 
with 249 regions 
Harvard-Oxford 
Atlas with 69 76 
regions 
MSDL Atlas with 
39 regions 
Power Atlas 
with 264 regions 
78.9&0.8 
77.6&1. 
79.3&0.7 
78.1&1. 
78.8&1. 
78.2&2.4 
75&1.5 
73.8&2.6 
72.4&1.4 
76.5&1.5 

*Fig \*\*\* Comparing performance of two atlas combinations used as inputs to a Dense FeedFwd model. The lower triangle outlined in purple also included structural information, while the blue box had only the combined 2-atlas featuresets. No combination outperformed the top singular models.*

***Ensembled Models***

Sets of top 10 models of selected atlas combinations were exhaustively combined in ensembles of 3 and 5 with either soft voting, hard voting, a logistic ridge regression ensembler, or a linear svm ensembler. The chosen sets to combine included the top 10 performing models of the Dense FeedFwd network trained on the combined structural and functional MRI parcellated with the BASC atlas with 122 ROIs and with 197 ROIs, the LSTM network using the fMRI data alone with the BASC atlas with 122 ROIs, and the Linear Ridge regression using combined structural and functional data and the BASC 122 atlas. This subset was chosen as these 5 categories were the highest performing models. *Table \*\*\** displays the performance of the top ensembles on the held-out test data. The top ensembles were chosen with 10x cross validation on the training set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Soft Voting** | **Hard Voting** | **Logistic Ridge** | **Linear SVM** |
| **Top model alone** | *0.804 Area under ROC curve* | | | |
| **Ensemble of 3** | 0.77609 | 0.76064 | 0.77730 | 0.75986 |
| **Ensemble of 5** | 0.77297 | 0.74681 | 0.75060 | 0.75565 |

*Table \*\*\*. Performance of the top ensembles on the held-out test data. The top ensembles were chosen with 10x cross validation on the training set, and their AUROC on the test set is displayed here.*

There were no gains in performance over using individual models, and due to the large increase in model complexity, the models were not examined in greater depth. The lack of increase in performance, coupled with this large random search and the results of the IMPAC challenge (cite) suggest that we may be approaching the limit of performance.

***Top Model Performance Ranges***

|  |  |
| --- | --- |
| ***Feature Set*** | ***Range of Performances (AUROC)*** |
| ***Combined Anatomical Volumetric Data and Connectivity with the BASC Atlas with 64 Parcellations*** | ***75.48-77.34*** |
| ***Combined Anatomical Volumetric Data and Connectivity with the BASC Atlas with 122 Parcellations*** | ***78.23-80.40*** |
| ***Combined Anatomical Volumetric Data and Connectivity with the BASC Atlas with 197 Parcellations*** | ***76.48-79.08*** |

*Table \*\*\* shows the performance of the set of top 5 models from the random search process. Due to the nature of holding out a single unbiased test set, error bars on the measures in Fig. X could not be generated. However, for a reasonable estimation of the ranges of performance our approach generated, we look at the top 5 models by 3-fold cross-validation and the range of performances encapsulated therein.*

***Un-tuned model performances:***

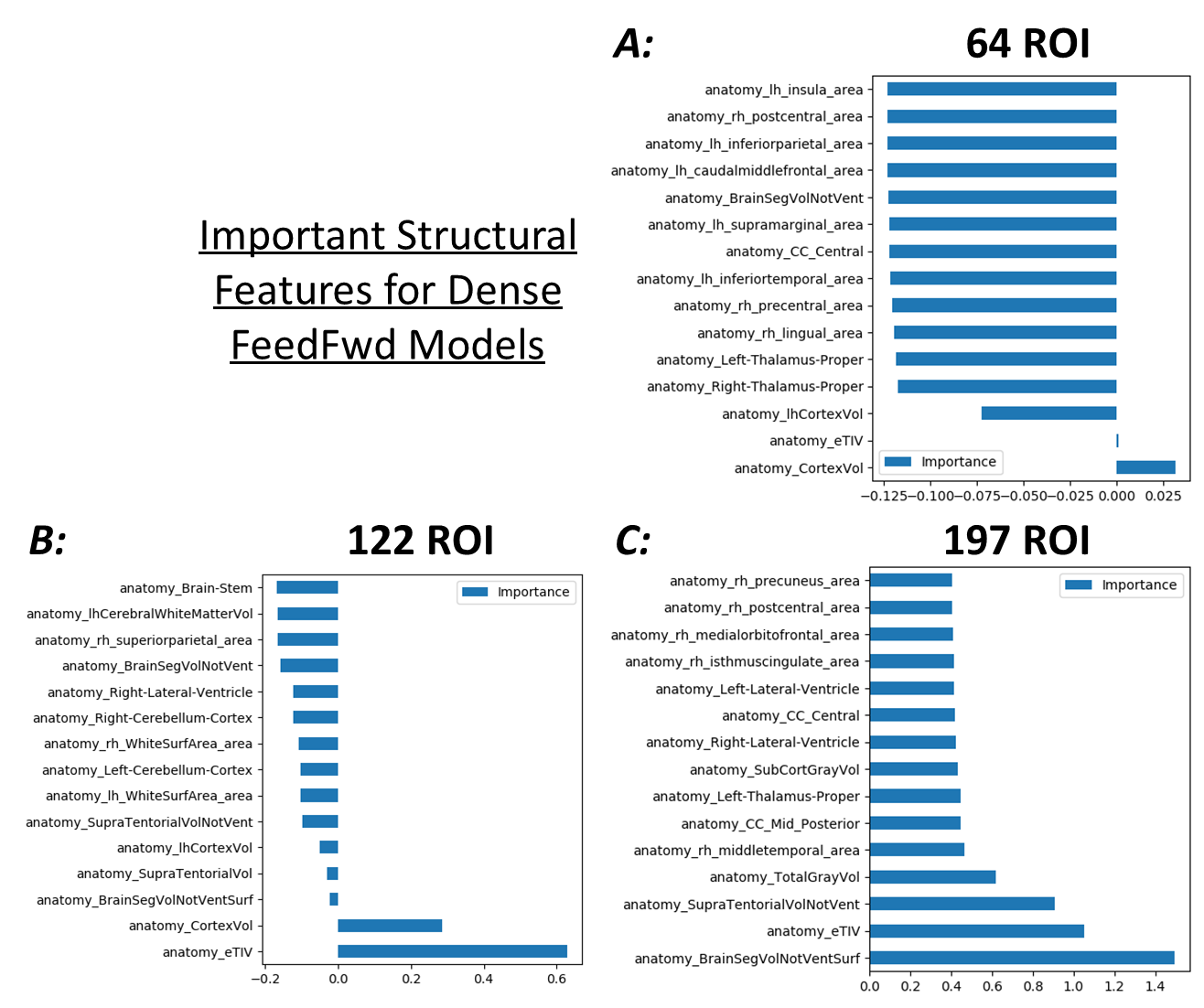
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Top Model on IMPAC | 2nd Best Model on IMPAC | 3rd Best Model on IMPAC | 4th Best Model on IMPAC | 5th Best Model on IMPAC |
| BASC atlas, 64 ROIs | 64.02 | 56.11 | 63.62 | 61.73 | 62.49 |
| BASC atlas, 122 ROIs | 65.38 | 63.98 | 54.13 | 67.64 | 66.13 |
| BASC atlas, 197 ROIs | 54.76 | 64.04 | 69.16 | 57.64 | 66.89 |

*Table S\*\*\*: AUROC of the top performing models on a novel, unseen dataset with no tuning or additional training. The models were tested on the ABIDEII dataset, NYU site.*

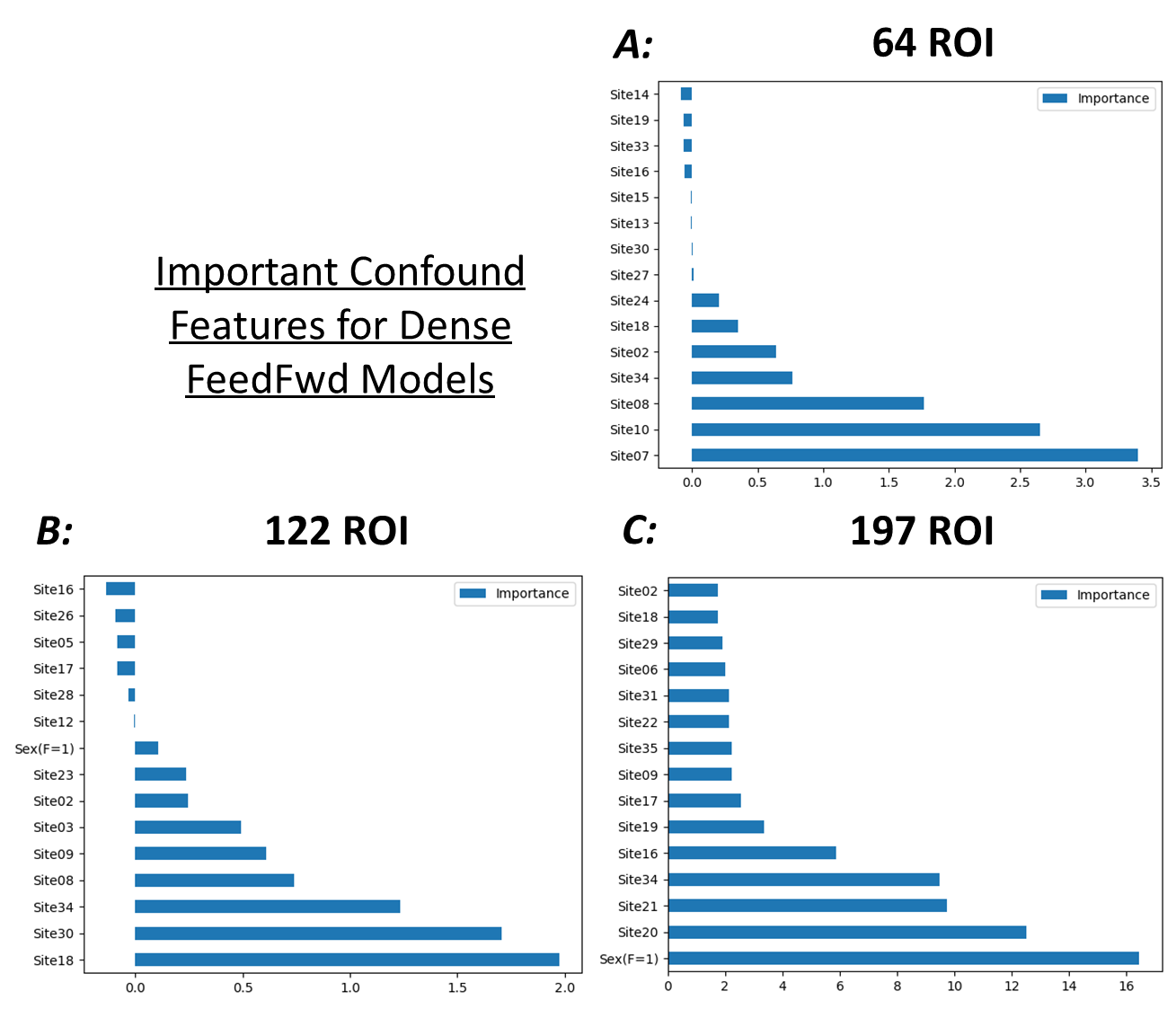
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Top Model on IMPAC | 2nd Best Model on IMPAC | 3rd Best Model on IMPAC | 4th Best Model on IMPAC | 5th Best Model on IMPAC |
| BASC atlas, 64 ROIs | 48.61 | 53.90 | 51.21 | 47.47 | 47.14 |
| BASC atlas, 122 ROIs | 48.13 | 50.78 | 50.23 | 47.74 | 49.12 |
| BASC atlas, 197 ROIs | 57.74 | 54.17 | 53.73 | 51.94 | 53.99 |

*Table S\*\*\*: AUROC of the top performing models on a novel, unseen dataset with no tuning or additional training. The models were tested on the ABIDEI dataset, NYU site.*

***Additional Feature Importance***

****

*Fig \*\*\* Structural Feature Importances: Sorted by z-score, average of the top 15 most important structural features (as found by permutation feature importance testing) from across the top 5 DenseFeedFwd models using the coarse atlas (A) medium atlas (B) and fine atlas (C).*

**

*Fig \*\*\* Confound feature Importances: Sorted by z-score, average of the top 15 most important confound features (as found by permutation feature importance testing) from across the top 5 DenseFeedFwd models using the coarse atlas (A) medium atlas (B) and fine atlas (C).*

References

[1] R. Toro, N. Traut, A. Beggatio, K. Heuer, and G. Varoquaux et al., *IMPAC: Imaging-psychiatry challenge: predicting autism: a data challenge on autism spectrum disorder detection*. Paris, 2018. [Online]. Available: https://​paris-saclay-cds.github.io​/​autism\_​challenge/​

[2] Kamalaker Dadi *et al.,* “Benchmarking functional connectome-based predictive models for resting-state fMRI,” *HAL-Inria*, 2018.

[3] Gael Varoquaux, “MRI Biomarkers Extraction: Teachings from an Autism-Prediction Challenge,” Granada, Spain, Sep. 20 2018.

[4] A. C. Stanfield, A. M. McIntosh, M. D. Spencer, R. Philip, S. Gaur, and S. M. Lawrie, “Towards a neuroanatomy of autism: a systematic review and meta-analysis of structural magnetic resonance imaging studies,” *European psychiatry : the journal of the Association of European Psychiatrists*, vol. 23, no. 4, pp. 289–299, 2008, doi: 10.1016/j.eurpsy.2007.05.006.

[5] A. Di Martino *et al.,* “The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism,” *Molecular psychiatry*, vol. 19, no. 6, pp. 659–667, 2014, doi: 10.1038/mp.2013.78.

[6] S. Parisot *et al.,* “Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease,” *Medical Image Analysis*, vol. 48, pp. 117–130, 2018, doi: 10.1016/j.media.2018.06.001.

[7] M. Khosla, K. Jamison, A. Kuceyeski, and M. Sabuncu, “3D Convolutional Neural Networks for Classification of Functional Connectomes,” Jun. 2018. [Online]. Available: http://​arxiv.org​/​pdf/​1806.04209v2

[8] C. Mellema, A. Treacher, K. Nguyen, and A. Montillo, “Multiple Deep Learning Architectures Achieve Superior Performance Diagnosing Autism Spectrum Disorder Using Features Extracted from Structural and Functional MRI,” *International Symposium on Biomedical Imaging*, 2019.

[9] S. Ghiassian, R. Greiner, P. Jin, and M. R. G. Brown, “Using Functional or Structural Magnetic Resonance Images and Personal Characteristic Data to Identify ADHD and Autism,” *PloS one*, vol. 11, no. 12, e0166934, 2016, doi: 10.1371/journal.pone.0166934.

[10] B. Sen, N. C. Borle, R. Greiner, and M. R. G. Brown, “A general prediction model for the detection of ADHD and Autism using structural and functional MRI,” *PloS one*, vol. 13, no. 4, e0194856, 2018, doi: 10.1371/journal.pone.0194856.

[11] H. Li, N. A. Parikh, and L. He, “A Novel Transfer Learning Approach to Enhance Deep Neural Network Classification of Brain Functional Connectomes,” *Frontiers in neuroscience*, vol. 12, p. 491, 2018, doi: 10.3389/fnins.2018.00491.

[12] J. Wang, Q. Wang, H. Zhang, J. Chen, S. Wang, and D. Shen, “Sparse Multiview Task-Centralized Ensemble Learning for ASD Diagnosis Based on Age- and Sex-Related Functional Connectivity Patterns,” *IEEE transactions on cybernetics*, vol. 49, no. 8, pp. 3141–3154, 2019, doi: 10.1109/TCYB.2018.2839693.

[13] T.-E. Kam, H.-I. Suk, and S.-W. Lee, “Multiple functional networks modeling for autism spectrum disorder diagnosis,” *Human brain mapping*, vol. 38, no. 11, pp. 5804–5821, 2017, doi: 10.1002/hbm.23769.

[14] A. Kazeminejad and R. C. Sotero, “Topological Properties of Resting-State fMRI Functional Networks Improve Machine Learning-Based Autism Classification,” *Frontiers in neuroscience*, vol. 12, p. 1018, 2018, doi: 10.3389/fnins.2018.01018.

[15] H. Huang, X. Liu, Y. Jin, S.-W. Lee, C.-Y. Wee, and D. Shen, “Enhancing the representation of functional connectivity networks by fusing multi-view information for autism spectrum disorder diagnosis,” *Human brain mapping*, vol. 40, no. 3, pp. 833–854, 2019, doi: 10.1002/hbm.24415.

[16] B. Yamagata *et al.,* “Machine learning approach to identify a resting-state functional connectivity pattern serving as an endophenotype of autism spectrum disorder,” *Brain imaging and behavior*, vol. 13, no. 6, pp. 1689–1698, 2019, doi: 10.1007/s11682-018-9973-2.

[17] O. Dekhil *et al.,* “Using resting state functional MRI to build a personalized autism diagnosis system,” *PloS one*, vol. 13, no. 10, e0206351, 2018, doi: 10.1371/journal.pone.0206351.

[18] R. S. Desikan *et al.,* “An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest,” *NeuroImage*, vol. 31, no. 3, pp. 968–980, 2006, doi: 10.1016/j.neuroimage.2006.01.021.

[19] P. Bellec, P. Rosa-Neto, O. C. Lyttelton, H. Benali, and A. C. Evans, “Multi-level bootstrap analysis of stable clusters in resting-state fMRI,” *NeuroImage*, vol. 51, no. 3, pp. 1126–1139, 2010, doi: 10.1016/j.neuroimage.2010.02.082.

[20] R. C. Craddock, G.A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, “A whole brain fMRI atlas generated via spatially constrained spectral clustering,” *Hum. Brain Mapp.*, vol. 33, no. 8, pp. 1914–1928, 2012, doi: 10.1002/hbm.21333.

[21] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion, “Multi-subject Dictionary Learning to Segment an Atlas of Brain Spontaneous Activity,” in *Lecture Notes in Computer Science, Information Processing in Medical Imaging*, G. Székely and H. K. Hahn, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 562–573.

[22] J. D. Power *et al.,* “Functional network organization of the human brain,” *Neuron*, vol. 72, no. 4, pp. 665–678, 2011, doi: 10.1016/j.neuron.2011.09.006.

[23] Gaël Varoquaux, Flore Baronnet, Andreas Kleinschmidt, Pierre Fillard, and and Bertrand Thirion, “Detection of Brain Functional-Connectivity Difference in Post-stroke Patients Using Group-Level Covariance Modeling,” *MICCAI*, 2010.

[24] J. Kawahara *et al.,* “BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment,” *NeuroImage*, vol. 146, pp. 1038–1049, 2017, doi: 10.1016/j.neuroimage.2016.09.046.

[25] W. Yan, H. Zhang, J. Sui, and D. Shen, “Deep Chronnectome Learning via Full Bidirectional Long Short-Term Memory Networks for MCI Diagnosis,” *Medical image computing and computer-assisted intervention : MICCAI … International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 11072, pp. 249–257, 2018, doi: 10.1007/978-3-030-00931-1\_29.

[26] Andrej Karpathy, *The Unreasonable Effectiveness of Recurrent Neural Networks.* [Online]. Available: http://​karpathy.github.io​/​2015/​05/​21/​rnn-​effectiveness/​

[27] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, “Permutation importance: a corrected feature importance measure,” *Bioinformatics (Oxford, England)*, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.

[28] Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux, “Extracting Universal Representations of Cognition across Brain-Imaging Studies,” *European psychiatry : the journal of the Association of European Psychiatrists*, 2018.

[29] M. Jung *et al.,* “Default mode network in young male adults with autism spectrum disorder: relationship with autism spectrum traits,” *Molecular autism*, vol. 5, p. 35, 2014, doi: 10.1186/2040-2392-5-35.

[30] M. Jung *et al.,* “Sex Differences in the Default Mode Network with Regard to Autism Spectrum Traits: A Resting State fMRI Study,” *PloS one*, vol. 10, no. 11, e0143126, 2015, doi: 10.1371/journal.pone.0143126.

[31] L. Q. Uddin *et al.,* “Salience network-based classification and prediction of symptom severity in children with autism,” *JAMA psychiatry*, vol. 70, no. 8, pp. 869–879, 2013, doi: 10.1001/jamapsychiatry.2013.104.

[32] C. P. Chen *et al.,* “Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default mode, and visual regions in autism,” *NeuroImage. Clinical*, vol. 8, pp. 238–245, 2015, doi: 10.1016/j.nicl.2015.04.002.

[33] A. Di Martino *et al.,* “Aberrant striatal functional connectivity in children with autism,” *Biological psychiatry*, vol. 69, no. 9, pp. 847–856, 2011, doi: 10.1016/j.biopsych.2010.10.029.

[34] J. D. Rudie *et al.,* “Autism-associated promoter variant in MET impacts functional and structural brain networks,” *Neuron*, vol. 75, no. 5, pp. 904–915, 2012, doi: 10.1016/j.neuron.2012.07.010.

[35] R. Bhaumik, A. Pradhan, S. Das, and D. K. Bhaumik, “Predicting Autism Spectrum Disorder Using Domain-Adaptive Cross-Site Evaluation,” *Neuroinformatics*, vol. 16, no. 2, pp. 197–205, 2018, doi: 10.1007/s12021-018-9366-0.