Brain Biomarker Interpretation in ASD Using Deep Learning and fMRI

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

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

Abstract. Autism spectrum disorder (ASD) is a complex neurodevelopmental disorder. Finding the biomarkers associated with ASD is extremely helpful to understand the underlying roots of the disorder and can lead to earlier diagnosis and more targeted treatment. Although Deep Neural Networks (DNNs) have been applied in functional magnetic resonance imaging (fMRI) to identify ASD, understanding the data driven computational decision making procedure has not been previously explored. Therefore, in this work, we address the problem of interpreting reliable biomarkers associated with identifying ASD; specifically, we propose a 2-stage method that classifies ASD and control subjects using fMRI images and interprets the saliency features activated by the classifier. First, we trained an accurate DNN classifier. Then, for detecting the biomarkers, different from the DNN visualization works in computer vision, we take advantage of the anatomical structure of brain fMRI and develop a frequency-normalized sampling method to corrupt images. Furthermore, in the ASD vs. control subjects classification scenario, we provide a new approach to detect and characterize important brain features into three categories. The biomarkers we found by the proposed method are robust and consistent with previous findings in the literature. We also validate the detected biomarkers by neurological function decoding and comparing with the DNN activation maps.

1 Introduction

Autism spectrum disorder (ASD) affects the structure and function of the brain. To better target the underlying roots of ASD for diagnosis and treatment, efforts to identify reliable biomarkers are growing [1]. Significant progress has been made using functional magnetic resonance imaging (fMRI) to characterize the brain changes that occur in ASD [2].

This work was supported by NIH Grant 5R01 NS035193.

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11072, pp. 206–214, 2018.
https://doi.org/10.1007/978-3-030-00931-1_24

Recently, many deep neural networks (DNNs) have been effective at identifying ASD using fMRI [3,4]. However, these methods lack model transparency. Despite promising results, the clinicians typically want to know if the model is trustable and how to interpret the results. Motivated by this, here we focus on developing the interpretation method for deciphering the regions in fMRI brain images that can distinguish ASD vs. control by the deep neural networks.

There are three main approaches for interpreting the important features detected by DNNs. One approach is using gradient ascent methods to generate an image that best represents the class [5]. However, this method cannot handle nonlinear DNNs well. The second approach is to visualize how the network responds to a specific corrupted input image in order to explain a particular classification made by the network [6]. The third one uses the intermediate outputs of the network to visualize the feature patterns [7]. However, all of these existing methods tend to end up with blurred and imprecise saliency maps.

The goal of our work is to identity biomarkers for ASD, defined as important regions of interest (ROIs) in the brain that distinguish autistic and healthy controls. Different from traditional brain biomarker detection methods, by utilizing the high dimensional feature capturing ability of DNNs and brain structure, we propose an innovative 2-stage pipeline to interpret biomarkers. Different from above DNN visualization methods, our main contribution includes a ROI-based image corruption and generating procedure. In addition, we analyze the feature importance using the distribution of DNN predictions and statistical hypothesis testing. We applied the proposed method on multiple datasets and validated our robust findings by decoding neurological function of biomarkers, viewing DNN intermediate outputs and comparing literature reports.

2 Method

2.1 Two-Stage Pipeline with Deep Neural Network Classifier

We propose a corrupting strategy to find the important regions activated by a well-trained ASD classifier (Fig. 1). The first stage is to train a DNN classifier for classifying ASD vs. control subjects. The DNN we use (2CC3D) has 6 convolutional, 4 max-pooling and 2 fully connected layers, followed by a sigmoid output layer [4] as shown in the middle of Fig. 1. The number of kernels are denoted on each layer in Fig. 1. Dropout and l2 regularization are applied to avoid overfitting. The study in [4] demonstrated that we can achieve higher accuracy using the 2CC3D framework, since it integrates spatial-temporal information of 4D fMRI. Each frame of 3D fMRI is downsampled to 32 $\times$ 32 $\times$ 32. We use sliding-windows with size w and stride length $stride$ to move along the time dimension of the 4D fMRI sequence and calculate the mean and standard deviation (std) for each voxel's time series within the sliding window. Given T frames in each 4D fMRI sequence, by this method, $\lfloor \frac{T-w}{stride} \rfloor + 1$ 2-channel images (mean and std fMRI images) are generated for each subject. We define the original fMRI sequence as $I(x, y, z, t)$, the mean-channel sequence as $\tilde{I}(x, y, z, t)$ and the std-channel as $\hat{I}(x, y, z, t)$. For any x, y, z in $\{0, 1, \cdots, 31\}$,

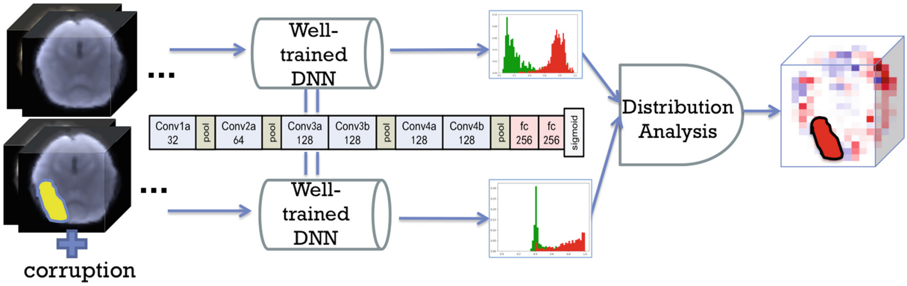


Fig. 1. Pipeline for interpreting important features from a DNN

Algorithm 1. Important Feature Detection For Binary Classification

Input: X^0, a group of images from class 0; X^1, a group of images from class 1; f, DNN classification model.

```
$\mathbb{P}^0_o \leftarrow f(X^0)$ and $\mathbb{P}^1_o \leftarrow f(X^1)$
$JSD^o_{+/-} \leftarrow JSD(\mathbb{P}^0_o, \mathbb{P}^1_o)$                                ▷ by bootstrapping
for $r$ in ROIs do
    $\mathbb{P}^0_c \leftarrow f(X^0_{\setminus r})$, $\mathbb{P}^1_c \leftarrow f(X^1_{\setminus r})$                        ▷ by sampling
    $JSD^c_{+/-} \leftarrow JSD(\mathbb{P}^0_c, \mathbb{P}^1_c)$, $Shift^0 \leftarrow \mathbb{P}^0_c - \mathbb{P}^0_o$, $Shift^1 \leftarrow \mathbb{P}^1_c - \mathbb{P}^1_0$
    if $JSD^c_+ < JSD^o_-$ or $median(\mathbb{P}^0_c) > median(\mathbb{P}^1_c)$ then     ▷ fool the classifier
        do Wilcoxon(Shift) one tailed test
        if $\mathbb{P}^0 \Rightarrow 1$ and $\mathbb{P}^1 \Rightarrow 0$ then
            r is an important feature for both classes
        else if only $\mathbb{P}^0 \Rightarrow 1$ then
            r is an important feature for class 0
        else if only $\mathbb{P}^1 \Rightarrow 0$ then
            r is an important feature for class 1
        end if                                              ▷ ⇒ means significant shift
    end if
end for
```

$$\tilde{I}(x, y, z, t) = \frac{\sum_{\tau=t+1-w}^{t} I(x, y, z, \tau)}{w} \tag{1}$$

$$\hat{I}(x,y,z,t) = \sqrt{\frac{\sum_{\tau=t+1-w}^{t} [I(x,y,z,\tau) - \tilde{I}(x,y,z,t)]^2}{w-1}}. \tag{2}$$

The outputs are probabilistic predictions ranging in $[0, 1]$. The second stage is to interpret the output differences after corrupting the image. We corrupt a ROI of the original image and put it in the well-trained DNN classifier to get a new prediction (Sect. 2.2). Based on the prediction difference, we use a statistical method to interpret the importance of the ROI (Sect. 2.3).

2.2 Prediction Difference Analysis

We use a heuristic method to estimate the feature (an image ROI) importance by analyzing the probability of the correct class predicted by the corrupted image.

In the DNN classifier case, the probability of the abnormal class c given the original image $\boldsymbol{X}$ is estimated from the predictive score of the DNN model f : $f(\boldsymbol{X}) = p(c|\boldsymbol{X})$. Denote the image corrupted at ROI $\boldsymbol{r}$ as $\boldsymbol{X}_{\setminus \boldsymbol{r}}$. The prediction of the corrupted image is $p(c|\boldsymbol{X}_{\setminus \boldsymbol{r}})$. To calculate $p(c|\boldsymbol{X}_{\setminus \boldsymbol{r}})$, we need to marginalize out the corrupted ROI $\boldsymbol{r}$:

$$p(c|\boldsymbol{X}_{\setminus \boldsymbol{r}}) = \mathbb{E}_{\boldsymbol{x_r} \sim p(\boldsymbol{x_r}|\boldsymbol{X}_{\setminus \boldsymbol{r}})} p(c|\boldsymbol{X}_{\setminus \boldsymbol{r}}, \boldsymbol{x_r}), \tag{3}$$

where $\boldsymbol{x_r}$ is a sample of ROI r. Modeling $p(\boldsymbol{x_r}|\boldsymbol{X}_{\setminus \boldsymbol{r}})$ by a generative model can be computationally intensive and may not be feasible. We assumed that an important ROI contains features that cannot be easily sampled from the same ROI of other classes and is predictive for predicting its own class. Hence, we approximated $p(\boldsymbol{x_r}|\boldsymbol{X}_{\setminus \boldsymbol{r}})$ by sampling $\boldsymbol{x_r}$ from each ROI $\boldsymbol{r}$ in the whole sample set. In fMRI study, each brain can be registered to the same atlas, so the same ROI in different images have the same spatial location and number of voxels. Therefore, we can directly sample $\hat{\boldsymbol{x}}_{\boldsymbol{r}}$s and replace $\boldsymbol{x_r}$ with them. Then we flatten the $\hat{\boldsymbol{x}}_{\boldsymbol{r}}$ and $\boldsymbol{x_r}$ as vectors $\overrightarrow{\hat{\boldsymbol{x}}_{\boldsymbol{r}}}$ and $\overrightarrow{\boldsymbol{x_r}}$. From the K sampled $\hat{\boldsymbol{x}}_{\boldsymbol{r}}^k$s, we calculate the Pearson correlation coefficient $\rho_k = cov(\overrightarrow{\hat{\boldsymbol{x}}_{\boldsymbol{r}}^{\boldsymbol{k}}}, \overrightarrow{\boldsymbol{x_r}})/\sigma_{\overrightarrow{\hat{\boldsymbol{x}}_{\boldsymbol{r}}^{k}}}\sigma_{\overrightarrow{\boldsymbol{x_r}}}$, where $k \in \{1, 2, \ldots, K\}$, $\rho \in [-1, 1]$. Because sample size of each class may be biased, we will de-emphasize the samples that can be easily sampled, since $p(c|\boldsymbol{X}_{\setminus \boldsymbol{r}})$ should be irrelevant to the sample set. Therefore, we will do a frequency-normalized transformation. We divide $[-1,1]$ into N equal-length intervals. Each ρ_k will fall in one of the intervals. After K samplings, we calculate N_i, the number of sample correlations in interval i, where $i \in \{1, 2, \ldots, N\}$. For the ρ_k located in interval i, the frequency-normalized weight is $w_k = \frac{1}{N \cdot N_i}$. Denote $\boldsymbol{X}'_{\boldsymbol{k}}$ as the image $\boldsymbol{X}$ replacing $\boldsymbol{x_r}$ with $\hat{\boldsymbol{x}}_{\boldsymbol{r}}^k$. Hence, we approximate $p(c|\boldsymbol{X}_{\setminus \boldsymbol{r}})$ as

$$p(c|\boldsymbol{X}_{\setminus \boldsymbol{r}}) \approx \sum_k w_k p(c|\boldsymbol{X}'_{\boldsymbol{k}}). \tag{4}$$

2.3 Important Feature Interpretation

In the binary classification scenario, we label the reference class as 0 and the experiment class as 1. The original prediction probability of the two classes are denoted as $\mathbb{P}_o^0$ and $\mathbb{P}_o^1$, which are two vectors containing the prediction results $p(c|\boldsymbol{X})$s for each sample in the two classes respectively. Similarly, we have $\mathbb{P}_c^0$ and $\mathbb{P}_c^1$ containing $p(c|\boldsymbol{X}_{\setminus r})$s for the corrupted images. We assume that corrupting an important feature will make the classifier perform worse. One extreme case is that the two distributions shift across each other, which can be approximately measured by $median(\mathbb{P}_c^0) > median(\mathbb{P}_c^1)$. If this is not the case, we use Jensen-Shannon Divergence (***JSD***) to measure the distance of two distributions:

$$JSD(\mathbb{P}_0, \mathbb{P}_1) = \frac{1}{2} KL(\mathbb{P}_0 \parallel \frac{\mathbb{P}_0 + \mathbb{P}_1}{2}) + \frac{1}{2} KL(\mathbb{P}_1 \parallel \frac{\mathbb{P}_0 + \mathbb{P}_1}{2}) \tag{5}$$

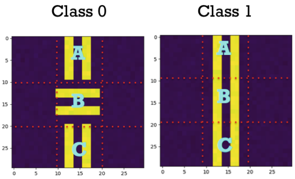


Fig. 2. Synthetic images

Table 1. Misclassification rate when corrupting patch B

	Class 0	Class 1
Equal	0.10 ± 0.01	0.91 ± 0.03
Normalize	0.49 ± 0.02	0.50 ± 0.01

where $KL(\mathbb{P}_0 \parallel \mathbb{P}_1) = -\sum_i \mathbb{P}_0(i) log(\mathbb{P}_1(i)/\mathbb{P}_0(i))$. Given two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, we use bootstrap method to calculate the upper bound JSD_+ and the lower bound JSD_- with confidence level, $1 - \alpha_{JSD}$. We classify the important ROIs into different categories based on the shift of the prediction distribution before and after corruption. The one-tailed Wilcoxon paired difference test [8] is applied to investigate whether the shift is significant. We use false discovery rate (FDR) controlling procedure to handle testing the large number of ROIs. FDR adjusted q-value is used to compare with the significance level α_W. The method to evaluate the feature importance is shown in Algorithm 1.

3 Experiments and Results

3.1 Experiment 1: Synthetic Data Model

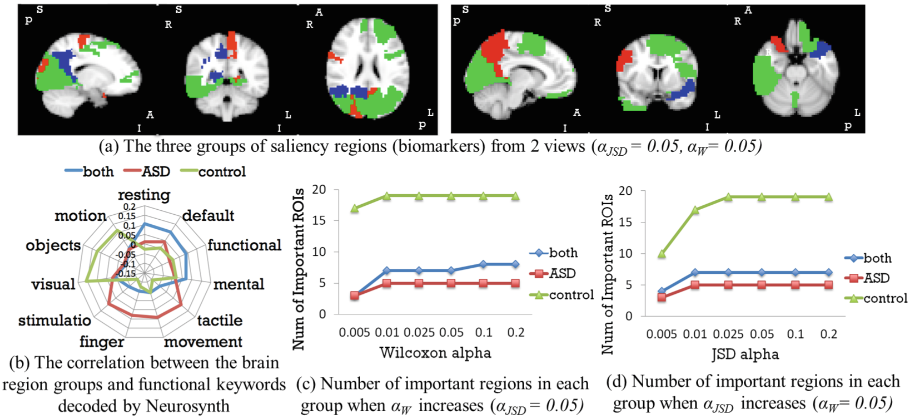
We used simulated experiments to demonstrate that our frequency-normalized resampling algorithm recovers the ground truth patch importance. We simulated two classes of images as shown in Fig. 2, with background = 0 and strips = 1 and Gaussian noise ($\mu = 0, \sigma = 0.01$). They can be gridded into 9 patches. We assumed that patch B of class 0 and 1 are **equally important** to human understanding. However, in our synthetic model, the sample set was biased with 900 images in class 0 and 100 images in class 1. A simple 2-layer convolutional neural network was used as the image classifier, which achieved 100% classification accuracy. Since the shift of corrupted images was obvious, we used misclassification rate to measure whether $p(c|\boldsymbol{X}_{\setminus r})$ was approximated reasonably by equally weighted sampling (which means $w_i = 1/K$) or by our frequency-normalized sampling. In Table 1, our frequency-normalized sampling approach ('Normalize') is superior to the equally weighted one ('Equal') in treating patch B equally in both classes.

3.2 Experiment 2: Task-fMRI Experiment

We tested our methods on a group of 82 ASD children and 48 age and IQ-matched healthy controls. Each subject underwent a task fMRI scan (BOLD, TR = 2000 ms, TE = 25 ms, flip angle = 60°, voxel size 3.44 × 3.44 × 4 mm^3) acquired on a Siemens MAGNETOM Trio TIM 3T scanner.

For the fMRI scans, subjects performed the "biopoint" task, viewed point light animations of coherent and scrambled biological motion in a block design

[2] (24 s per block). The fMRI data was preprocessed using FSL [9] for (1) motion correction, (2) interleaved slice timing correction, (3) BET brain extraction, (4) spatial smoothing (FWHM = 5 mm), and (5) high-pass temporal filtering. The functional and anatomical data were registered and parcellated by AAL atlas [10] resulting in 116 ROIs. We applied a sliding window ($w = 3$) along the time dimension of the 4D fMRI, generating 144 3D volume pairs (mean and std) for each subject.



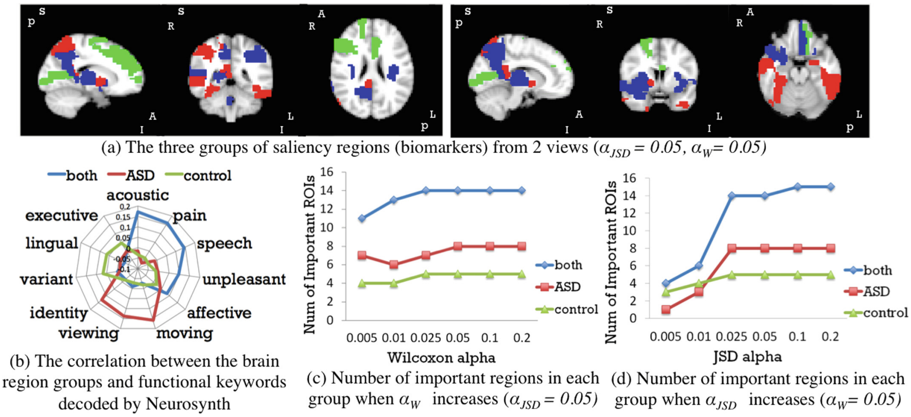
(a) The three groups of saliency regions (biomarkers) from 2 views ($\alpha_{JSD} = 0.05, \alpha_W = 0.05$)

(b) The correlation between the brain region groups and functional keywords decoded by Neurosynth

(c) Number of important regions in each group when α_W increases ($\alpha_{JSD} = 0.05$)

(d) Number of important regions in each group when α_{JSD} increases ($\alpha_W = 0.05$)

Fig. 3. Important biomarkers detected in biopoint dataset



(a) The three groups of saliency regions (biomarkers) from 2 views ($\alpha_{JSD} = 0.05, \alpha_W = 0.05$)

(b) The correlation between the brain region groups and functional keywords decoded by Neurosynth

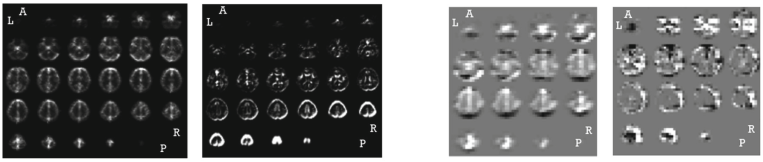
(c) Number of important regions in each group when α_W increases ($\alpha_{JSD} = 0.05$)

(d) Number of important regions in each group when α_{JSD} increases ($\alpha_W = 0.05$)

Fig. 4. Important biomarkers detected in ABIDE dataset

We split 85% subjects (around 16k 3D volume pairs) as training set, 7% as validation set for early stopping and 8% as testing set, stratified by class.

The model achieved 87.1% accuracy when evaluated on each 3D pair input of the testing set. Figure 3(a) and (b) give two views of the important ROIs brain map ($\alpha_{JSD} = 0.05$, $\alpha_W = 0.05$). Blue ROIs are associated with identifying both ASD and control. Red ROIs are associated with identifying ASD only and green ROIs are associated with identifying control only. By decoding the neurological functions of the important ROIs with Neurosynth [11], we found (1) regions related to default mode and functional connectivity are significant in classifying both individuals with ASD and controls, which is consistent with prior literature related to executive functioning and problem-solving in ASD [2]; (2) regions associated with finger movement are relevant in classifying individuals with ASD, and (3) visual regions were involved in classifying controls, perhaps because controls may attend to the visual features more closely, whereas ASD subjects tend to count the dots on the video [12].



(a) ASD group outputs from Conv1a
Left: Biopoint Dataset Right: ABIDE Dataset

(b) Difference of outputs from Conv2a (Control-ASD)
Left: Biopoint Dataset Right: ABIDE Dataset

Fig. 5. Intermediate outputs (activation maps) of DNN

3.3 Experiment 3: Resting-State fMRI

We also performed experiments on data from the ABIDE I cohort UM site [9,13].This resulted in 41 ASD subjects and 54 healthy controls. Each subject initially had 293 frames. As in the task-fMRI experiment, we generated 2-channel images. We used the weights of the pre-trained 2CC3D networks in experiment 2 as our initial network weights. We split 33 ASD subjects and 43 controls for training (around 22k 3D volume pairs). 9 subjects were used as validation data for early stopping. The classifier achieved 85.3% accuracy in identifying individual 3D volume on the 10 subjects testing set. The biomarker detection results are shown in Fig. 4: (1) emotion related regions colored in blue are highlighted for both groups; (2) regions colored in red (viewing and moving related) are associated with identifying ASD; and (3) green regions (related to executive and lingual) are associated with identifying control.

3.4 Results Analysis

In experiment 2, since the subjects were under visual task, visual patterns were detected. Whereas in experiment 3, subjects were in resting state, so no visual regions were detected. In addition, we found many common ROIs in both experiments: frontal (motivation related), precuneus (execution related), etc. Previous research [2] also indicated these regions are associated with identifying ASD vs. control. Moreover, from the sub-figure (c), (d) of Figs. 3 and 4, the groups of detected important regions are very stable when tuning JSD confidence level (1-α_{JSD}) and Wilcoxon testing threshold α_W, except when α_{JSD} is very small. This is likely because the original prediction distribution is fat tailed. Furthermore, we validate the results with the activation maps from the 1st and 2nd layers of the DNN. The output of each filter was averaged for 10 controls and for 10 ASD subjects. The 1st convolutional layer captured structural information and distinguished gray vs. white matter (Fig. 5(a)). Its outputs are similar in both control and ASD group. The outputs of the 2nd convolutional layer showed significant differences between groups in Fig. 5(b). Regions darkened and highlighted in Fig. 5(b) correspond to many regions detected by our proposed method.

4 Conclusions

We designed a 2-stage (DNN + prediction distribution analysis) pipeline to detect brain region saliency for identifying ASD and control subjects. Our sampling and significance testing scheme along with the accurate DNN classifier ensure reliable biomarker detection results. Our method was designed for interpreting important ROIs for registered images, since the traditional machine learning feature selection methods can not be directly used in interpreting DNNs. Moreover, our proposed method can be directly used to interpret any other machine learning classifiers. Overall, the proposed method provides an efficient and objective way of interpreting the deep learning model applied to neuroimages.

References

1. Goldani, A.A., et al.: Biomarkers in autism. Front. Psychiatry **5**, 100 (2014)
2. Kaiser, M.D., et al.: Neural signatures of autism. In: PNAS (2010)
3. Iidaka, T.: Resting state functional magnetic resonance imaging and neural network classified autism and control. Cortex **63**, 55–67 (2015)
4. Li, X., et al.: 2-channel convolutional 3D deep neural network (2CC3D) for fMRI analysis: ASD classification and feature learning. In: ISBI (2018)
5. Yosinski, J., et al.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
6. Zintgraf, L.M., et al.: Visualizing deep neural network decisions: prediction difference analysis. arXiv preprint arXiv:1702.04595 (2017)
7. Zhou, B., et al.: Learning deep features for discriminative localization. In: CVPR. IEEE (2016)
8. Whitley, E., et al.: Statistics review 6: nonparametric methods. Crit. Care **6**, 509 (2002)
9. Smith, S.M., et al.: Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage **23**, S208–S219 (2004)
10. Tzourio-Mazoyer, N., et al.: Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage **15**, 273–289 (2002)

11. Yarkoni, T., et al.: Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods **8**, 665 (2011)
12. Ventola, P., et al.: Differentiating between autism spectrum disorders and other developmental disabilities in children who failed a screening instrument for ASD. J. Autism Dev. Disord. **37**, 425–436 (2007)
13. Di Martino, A., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry **19**, 659 (2014)