This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016734, IEEE Access

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Exploring the structural and strategic bases of autism spectrum disorders with deep learning

FENGKAI KE[1], SEUNG JIN CHOI[2], YOUNG HO KANG[3], KEUN-AH CHEON[2*], AND SANG WAN LEE[4*]

[1]Hubei Key Laboratory of Modern Manufacturing Quality Engineering, School of Mechanical Engineering, Hubei University of Technology, Wuhan, P.R.China (e-mail: kfkhbut@163.com)
[2]Division of Child and Adolescent Psychiatry, Department of Psychiatry, Severance Children's Hospital, Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, South Korea
[3]Brain and Cognitive Engineering Program, Korea Advanced Institute of Science Technology (KAIST), Daejeon 34141, Republic of Korea
[4]Brain and Cognitive Engineering Program, Department of Bio and Brain Engineering, Center for Neuroscience-inspired AI, KI for Health Science Technology, KI for Artificial Intelligence, Korea Advanced Institute of Science Technology (KAIST), Daejeon 34141, Republic of Korea

Corresponding author: Keun-Ah Cheon (e-mail: KACHEON@yuhs.ac), Sang Wan Lee (e-mail: sangwan@kaist.ac.kr).

This work was supported in part by the Scientific and Technological Research of Education Department of Hubei Province under Grant Q20181408, in part by the Scientific Research Foundation of Science and Technology Department of Hubei Province under Grant 2018CFB276, in part by the Doctor Launching Fund of Hubei University of Technology under Grant BSQD20160004, and in part by the Hubei Chenguang Talented Youth Development Foundation (HBCG) under Grant 2017109. This work was also supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government (MSIT) (No.2019-0-01371, Development of brain-inspired AI with human-like intelligence), Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TC1603-06, Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) grant funded by the Korea government (MOHW) (HI12C0021-A120029) and in part by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2016R1A2B4006737).

ABSTRACT Deep learning models are applied in clinical research in order to diagnose disease. However, diagnosing autism spectrum disorders (ASD) remains challenging due to its complex psychiatric symptoms as well as a generally insufficient amount of neurobiological evidence. We investigated the structural and strategic bases of ASD using 14 different types of models, including convolutional and recurrent neural networks. Using an open source autism dataset consisting of more than 1000 MRI scan images and a high-resolution structural MRI dataset, we demonstrated how deep neural networks could be used as tools for diagnosing and analyzing psychiatric disorders. We trained 3D convolutional neural networks to visualize combinations of brain regions, thus representing the most referred-to regions used by the model whilst classifying the images. We also implemented recurrent neural networks to classify the sequence of brain regions efficiently. We found emphatic structural and strategic evidence on which the model heavily relies during the classification process. For instance, we observed that the structural and strategic evidence tends to be associated with subcortical structures, including the basal ganglia (BG). Our work identifies the distinct brain structures that characterize a complex psychiatric disorder while streamlining the deductive reasoning that clinicians can use to ensure an economical and time-efficient diagnosis process.

INDEX TERMS Deep Learning, sMRI, Austism Spectrum Disorders, Neural Networks

I. INTRODUCTION

AUTISM spectrum disorders (ASD) is a term embodying neurodevelopmental disorders characterized by persistent insufficiencies in social communication as well as restricted and repetitive behaviors, interests, or activities [1]. According to a report from the Centers for Disease Control and Prevention (CDC) in 2018 [2], one out of 59 children in the United States has ASD symptoms. In the Republic of Korea, the prevalence of ASD is estimated to be 2.64% among school-age children [3].

Studies using neuroimaging techniques, such as magnetic resonance imaging (MRI) or positron emission tomography (PET), have provided many insights into the neurodevelopmental characteristics underlying ASD [4]–[8]. Most findings from these imaging studies are based on a univariate analytical approach assuming the independence of each voxel [9] [10]. In contrast to mass-univariate methods, machine learning models can use multiple voxels as inputs, making it possible to study high-level relationships between different features. These models are capable of identifying the differences between a disease and control group, while suggesting a suitable diagnosis strategy for each subject [11]. Machine learning models have been successful in solving various disease classification problems in ailments including Alzheimer's disease [12]–[15], schizophrenia [16] [17], attention deficit hyperactivity disorder [18] [19], and other psychiatric diseases [20] [21].

II. RELATED WORK

Rapid advances in deep learning have allowed the integration of various data, including data with different modalities [22]–[27]. Several studies have demonstrated the utility of deep learning in medical problems [28]. For example, a fusion of latent feature representations extracted from MRI and PET data has been used in diagnosing Alzheimer's disease [29] [30]. Deep learning has performed well in learning complex patterns, such as functional connectivity, making it potentially helpful for diagnosis purposes [31].

ASD is characterized by persistent deficits in social communication and interaction as well as restricted and repetitive patterns of behavior, interests, or activities. The causes of ASD are still unknown, but some researchers hypothesize that the structure of the brain contains relevant information [32] [33]. The data consist of volumetric measures and the structure of the cerebellar vermis [34], regional thicknesses extracted from the surface-based morphometry [35], the volumes of gray and white matter maps [36] [37], volumetric and geometric features extracted from selected cortical locations, and morphometric features of selected regions of interest [38]. A few studies have reported a relatively high accuracy, between 76% and 90%. However, these studies involved performance measurement of classifiers conducted on small datasets, usually consisting of less than 50 participants [39]–[41]. Moreover, the body of research has yet to produce robust algorithms or out-of-sample performance.

When these tests were implemented on a large-scale dataset collected from different populations and places, their performance significantly decreased. One study used MRI samples from the Autism Brain Imaging Data Exchange (ABIDE) to define the histogram of oriented gradients, obtaining an accuracy of 60.1% [42]. In another study, two different types of neural networks were used to process MRI data. This study achieved an accuracy of 61.7%. The models reportedly performed better on relatively large-scale MRI datasets [43]. Weights from the convolution neural networks (CNN) were replaced with weights from the pre-trained sparse autoencoder network. In addition to inadequate classification performance, the models carry poor transparency. In other words, the factors affecting the model's decisions remain ambiguous whilst classifying each subject. Such factors can be used as indices measuring model suitability. After preprocessing 1113 sMRI samples from the Human Connectome Project (HCP) data set (http://www.humanconnectome.org/ for technical information) and screening out sMRI data with similar age and gender ratios in the ABIDE, we used an encoder to classify subjects with autism [44] [45]. This model can predict the neuroanatomical deviations associated with autism compared to a control group [46].

In addition to the sMRI-based classification, other studies also have used the fMRI data. Based on Pearson's coefficient, 19900 Region of Interest (ROI) features were selected from the CC200 functional parcellation atlas of the brain [47], and an autoencoder was used to classify autism, with an accuracy of 0.743. Similarly, by using the parcellation atlas [48], the temporal features of each ROI in the rs-fMRI data were calculated and fed into the 1D convolutional neural network, leading to a classification accuracy of 81% for the ABIDE-ETH1 dataset [49]. Another study employed a cross-validation grid search method was used to compare multiple classification models such as support vector machines, logistics and ridge regression. The classification accuracy was 71.98%. Researchers further analyzed the seven different brain atlas CC400 to identify autism correlated and anti-correlated region of interests in the brain [50].

Model comprehensibility is particularly crucial in diagnosing psychiatric diseases, especially when the causes of the disease are not fully known. Finding solutions to these fundamental issues is a necessity for enhancing both the reliability of classification performance and interpretability of the model's decision.

III. EXPLORING THE STRUCTURAL AND STRATEGIC BASES OF AUTISM SPECTRUM DISORDERS

To resolve these issues, we conducted large-scale simulations comparing the classification performance of five different categories of deep learning models, including convolutional neural networks, recurrent neural networks, and spatial transformation networks. We were able to visualize the results of each model. The simulations were carried out on two different neuroimaging datasets: one is from the Child Psychiatric Clinic at Severance Hospital, Yonsei University College of Medicine (YUM) which had a high-resolution structural MRI and another is from the international Autism Brain Imaging Data Exchange (ABIDE).

First, we carried out an extensive model comparison for reliable performance evaluation between a number of classifiers using various network architectures. Second, we explored the structural bases of ASD by visualizing a combination of brain regions, which can be considered the bases of the model's classification decision. Further, we included invariant classifiers in our study to effectively deal with variations in size and translation. Our findings suggest the possibility that ASD patients have distinctive structural signatures in their brains. Last but not least, we used attention-based recurrent neural networks to learn a sequence of the brain regions, leading to classification. This sequence provides a better understanding of the background strategies used by the models while classifying the data. Revealing such strategies pointed to regions of the brain for assessment when making a diagnosis. These strategies can make the diagnosis process more economical and time-efficient by providing a useful order of the regions associated with diseases. We observed compelling brain regions for the model's classification, particularly multiple subcortical structures, including the basal ganglia. Overall, these results provide both structural and strategic information for characterizing ASD, as shown in Fig. 1.

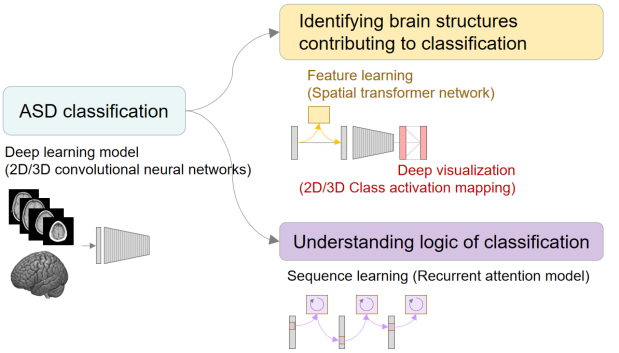

Identifying brain structures contributing to classification
Feature learning (Spatial transformer network)
ASD classification
Deep learning model (2D/3D convolutional neural networks)
Deep visualization (2D/3D Class activation mapping)
Understanding logic of classification
Sequence learning (Recurrent attention model)

FIGURE 1. Overall framework. Our framework is two-fold: one for learning to classify ASD (corresponding to the section: Training various types of neural networks for ASD classification); the other for visualizing the structural and logical basis of classification (corresponding to the two sections: Identifying brain structures contributing to the classification, Understanding the logic of classification).

IV. METHODS

A. DATA AND PRE-PROCESSING

In our study, we used two MRI datasets for autism classification research, the first collected by the Yonsei University College of Medicine (YUM). The second dataset was obtained from the Autism Brain Imaging Data Exchange (ABIDE) website, which houses a large number of open-source MRIs for autism research [51].

For the YUM dataset, according to the sample image quality, we selected 73 out of 84 samples, including 40 people with high SCQ points and 33 people with low SCQ points. All subjects gave informed consent, and the Institutional Review Board of the Severance Hospital of Yonsei University approved the study for research with human subjects. We performed this study at the Yonsei University College of Medicine. In addition, we confirm that all methods were performed in accordance with the relevant guidelines and regulations.

For the ABIDE dataset, after combining the ABIDE I and ABIDE II databases and screening the MRI data for suboptimal quality, there were 1,992 people in total, with 946 autism patients and 1,046 people as controls.

The ABIDE dataset is a combination of sets of MRI scans taken independently by more than 24 organizations, leading to inconsistency in MRI quality and dimensions. As a result, the dataset required cautious preprocessing.

For the YUM dataset, the processing pipeline consists of three steps: (A) transformation of the MRIs into the same size (170×256×256); (B) resizing of each image to a smaller size (85×128×128) for faster computation; and (C) normalization of the voxel values to a range of [0,1].

The pre-processing method employed for the ABIDE dataset differs from the YUM dataset. Because of the dissimilar configuration and quality of each dataset, we employed Statistical Parametric Mapping software (SPM8) to perform the registration [52]. The ABIDE pre-processing pipeline consisted of two steps: (A) non-linear spatial transformation of the MRI to the Montreal Neurological Institute (MNI) T1 template [53]; and (B) normalization of the voxel value to a range of [0,1]. In step (A), we used the default setting of the bounding box, which was [-78, -112, -50] to [78, 76, 85], and the voxel size, which is 2 mm×2 mm×2 mm, in the SPM8. The size of the MRI after registration became 79×95×79.

B. MODELS

In this paper, we used five main model configurations for classifying and visualizing the samples, as shown in Table 1. Some of them have several model subtypes. For example, we can use the 2D CNN or 3D CNN to process 2D MRI or 3D MRI input.

For model type 1, there are two subtypes: 3D input+3D CNN (1-1) and 3D input+2D CNN (1-2). We use the 3D MRI scan as input and the traditional 2D and 3D CNN for extracting the feature map and classification [26] [55].

We illustrate the architecture of model type 2 in Fig. 2A. The models combine the use of STN into the traditional CNN to look at the specific part of the MRI. There are also four model subtypes, which are 2D input+3D CNN+2D STN (2-1), 3D input+3D CNN+3D STN (2-2), 3D input+2D CNN+3D STN (2-3), 2D input+3D CNN+2D STN (2-4). Because we had to deal with the 3D input data, we modified the original the STN model to the 3D version of the STN [54], so-called 3D STN. That is, the 3D STN receives the three-dimensional input, and the spatial transformation matrix τ has been changed to 4×4, as follows

$$\tau = \begin{bmatrix} s_x & 0 & 0 & t_x \\ 0 & s_y & 0 & t_y \\ 0 & 0 & s_z & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

where s_x, s_y, s_z are the scale factors for each dimension and t_x, t_y, t_z are the translations for each dimension.

We have depicted the architecture of model type 3 in Fig. 2B. There are two types of models: 3D input+2D CNN+3D STN+RNN (3-1) and 3D input+3D CNN+3D STN+RNN (3-2). The architecture for model type 4 is shown in Fig. 2C - Fig. 2D. There are two types of model: 2D input+2D CNN+CAM (4-1) and 3D input+3D CNN+CAM (4-2). The core idea of CAM is to use the global averaging pooling (GAP) layer, $F^k = \sum f_k(x, y)$ for every (x, y) in order to calculate the importance of each slice of the feature map from

TABLE 1. Different Model Types

Type	Model	Model Subtype
1	3D input+2D/3D CNN	3D input+2D CNN (1-1) 3D input+3D CNN (1-2)
2	2D/3D input+2D/3D CNN+2D/3D STN	2D input+2D CNN+2D STN (2-1) 2D input+3D CNN+2D STN (2-2) 3D input+2D CNN+3D STN (2-3) 3D input+3D CNN+3D STN (2-4)
3	3D input+2D/3D CNN+3D STN+RNN	3D input+2D CNN+3D STN+RNN (3-1) 3D input+3D CNN+3D STN+RNN (3-2)
4	2D/3D input+2D/3D CNN+CAM	2D input+2D CNN+CAM (4-1) 3D input+3D CNN+CAM (4-2)
5	3D input+RAM	3D input+RAM+loc (5-1) 3D input+RAM+noloc (5-2) 3D input+RAM+loc+fc (5-3) 3D input+RAM+rand (5-4)

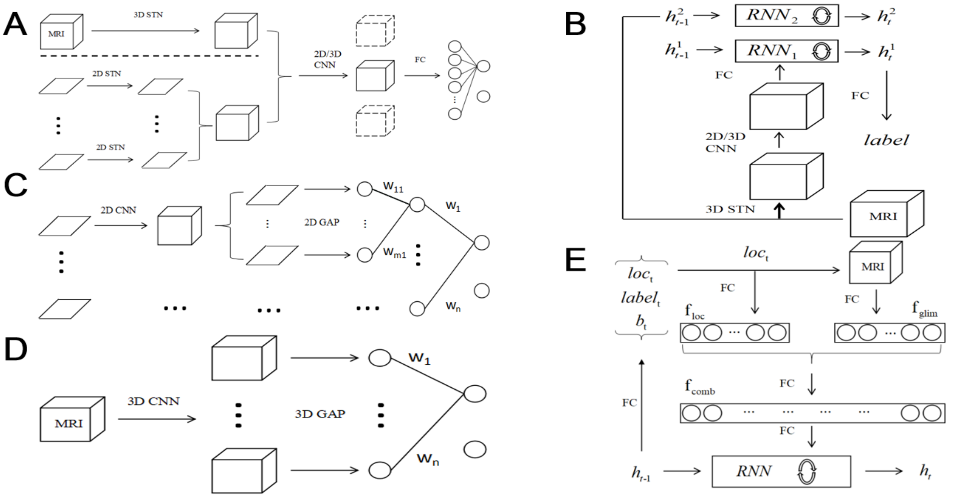


FIGURE 2. Network architectures for model type 2 to 5. The 3D cube and parallelogram represent the original MRI data and MRI slices, respectively. (A) Model type 2. There are two kinds of MRI inputs separated by the dashed line and two methods of convolution. The cube in solid lines and dash lines is the feature map extracted after 2D and 3D convolution layers, respectively. (B) Model type 3. The 3D STN here is slightly different from (A). For each time step in the RNN, the previously hidden state h2 in the second layer becomes the input of the STN to output the spatial transformation matrix. Then the STN uses it to transform the original 3D MRI image spatially [54]. (C) In model subtype 4-1, for each slice of the original MRI, their processing pipeline are the same and independent. The slice is input into a 2D CNN and becomes the 3D feature map. Then we use the 2D GAP to process each slice of this 3D feature map and fully connected to a single unit. Then all these single units from each slice are fully connected to the last layer for classification. (D) In model subtype 4-2, the MRI is fed into the 3D CNN, and we use 3D GAP to process each cube after the convolution. (E) Model type 5. For each time step of the RNN, the previous hidden state is fed into an FC layer, called the location network, to output the attention location. We use this location to extract the cubic patch, which is called the glimpse network. Then we use the FC layer to process the location and cubic patch to get location and glimpse features respectively and combine them.

the last convolution layer before creating a heat map for a given image using

$$M_c(x,y) = \sum_k w_k^c f_k(x,y) \tag{2}$$

where c stands for class, $f_k(x,y)$ is the activation value of k^{th} unit in the last convolution layer at the specific point (x,y), w_k^c stands for the weights of the FC layer that connects the unit in the GAP layer with the k^{th} unit in the output layer [56]. Because we used (2) for 2D images, we can call the (2) the 2D GAP. For 2D input+2D CNN+CAM (4-1), we use the 2D GAP layer for the feature map of each slice's last convolution layer in the MRI image $F_{ij}^k = \sum f_i^k(x,y)$ for every (x,y) and the (2) has been changed to

$$M_{i,c}(x,y) = \sum_k w_i^c w_{ik}^c f_i^k(x,y) \tag{3}$$

where i is the i^{th} MRI slice, and w_k^c and w_{ik}^c stands for the weights in the last two FC layers. The remaining symbols are the same as in (2). For 3D input+3D CNN+CAM (4-2), only the feature map's dimension after convolution has been changed to 4-dimensional, so (2) becomes

$$M_c(x,y,z) = \sum_k w_k^c f_k(x,y,z) \tag{4}$$

where $f_k(x,y,z)$ is the feature value of k^{th} unit in the last convolution layer at the specific point (x,y,z).

For model type 5, we show the architecture in Fig. 2E.

We change the traditional recurrent attention model from 2-dimensional image input to 3-dimensional MRI input [57]. We set the center of the MRI image to be the starting point of the RAM model. Initially, RAM used the last hidden state of the RNN for classification and did not have the location constraint in the cost function. From the experimental results, the location network inside the RAM always outputs the coordinates near the corner of the MRI, which means it converges to the local minima quickly. Thus, we added a constraint function (5) into the cost function in order to assist the RAM in learning more useful information and reaching the global optimum.

$$f(x,y,z) = \begin{cases} 0 & 0.2 < (x,y,z) < 0.8 \\ C & otherwise \end{cases} \tag{5}$$

where C is a constant value. The location (x, y, z) in the image has been normalized to the range of [0,1], with (0,0,0) being the top left corner of the image and (1,1,1) being the bottom right corner of the image. The equation above forces the RAM to focus on the central part of the brain. If not, it will be challenged by a constant value, C.

C. IMPLEMENTATION DETAILS

For training and testing data, we separated the training and testing data to be 80% and 20% of the original database. We made the percentage of patients with autism in the training data the same as in the original database. For each type of model, we used the 10-fold cross-validation method.

For hardware configuration, we primarily used an Intel Core i76700 CPU @ 3.40GHz×8 processor and a TITAN Xp/PCIe/SSE2 graphics processing unit.

We used the network architectures shown in Table 2.

$2DCNN(f_h/f_w, ks, s)$ is the abstraction of the 2-dimensional convolution layer with f_h number of filters for the YUM dataset and f_w for the ABIDE dataset, ks is the kernel size, and s is the stride. If f_w is not specified, it means the YUM and the ABIDE datasets share the same number of filters. $3DCNN(f, ks, s)$ follows a similar definition.

$2DMP(ps, s)$ is the abstraction of the 2-dimensional max-pooling layers withpool size and stride. $3DMP(ps, s)$ holds a similar definition.

$BATCH()$ is the abstraction of batch normalization, while $DROP(p)$ is the abstraction of the dropout layer with p probability. $FC(k)$ is the abstraction of a fully connected layer with a k output unit. $RNN(k)$ is the abstraction of the recurrent neural network with k output unit, $2DGAP()$ is the abstraction of the global averaging pooling layer, so as $2DGAP()$ for different dimensions.

For model type 1, the 3D input+2D CNN (1-1) and 3D input+3D CNN (1-2) models are shown in Table 2.

For model type 2, the 2D input+2D CNN+2D STN (2-1), 2D input+3D CNN+2D STN (2-2), 3D input+2D CNN+3D STN (2-3), and 2D input+3D CNN+2D STN (2-4) subtypes of the model are shown in Table 2. $N \times \{2DSTN()\}$ stands for concatenating N slice of the transformed MRI image along the first dimension after using the 2D STN model.

For model type 3, the 3D input+2D CNN+3D STN+RNN (3-1), 3D input, 3D CNN+3D STN+RNN (3-2) are shown in Table 2. We used the $RNN_1(128)$ and $RNN_2(128)$ to represent the two-layer RNN, the layers with ⋇ superscript are used recurrently in RNN.

For model type 4, the 2D input+2D CNN+CAM (4-1) and 3D input+3D CNN+CAM (4-2) are shown in Table 2, where the ⋇ superscript indicates that these layers are used for each slice of the MRI image repeatedly and independently. We used the central part of the original images as an input for the models.

Model type 5, it is rather awkward to summarize simply using a table. We give the implementation details of each network as described in [10]. At each time step, the glimpse network extracts three cubic patches inside the MRI image, with the size of the first cubic patch being $4 \times 4 \times 4$, and each successive patch having twice the width, height, and depth of the previous. After extracting and resizing them to the same size, we flattened the three cubic patches and inputted them into the fully connected layer with 128 units. The location network takes the location coordinate as input to the fully connected layer with 128 units. We then concatenated the glimpse feature from the glimpse network and the location feature from the location network into the combined feature, inserting them into the RNN with 256 units. After eight-time steps or glimpses, the hidden states of the RNN were used for classification. For 3D input+RAM+loc (5-1), the location constraint cost function is adopted inside the model. For 3D input+RAM+noloc (5-2), no location constraint cost function is used inside the model. For 3D input+RAM+loc+fc (5-3), the location constraint cost function is exploited inside the model and uses all the hidden states information for classification, while omitting the others. We set the center of the MRI image to be the starting location of the RAM model (5-1) to (5-3). For 3D input+RAM+rand (5-4), the location network of a Gaussian distribution function centered at 0 is replaced with a 0.6 standard deviation within the model. We found that, even with the RAM+noloc model, we could still reach a relatively high accuracy as RAM+loc, implying that the attention regions after the first time step of RAM are meaningless.

D. DATA AVAILABILITY

The ABIDE dataset analyzed during the current study is publicly available on http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html. Moreover, the YUM dataset that supports the findings of this study are available from Severance Children's Hospital, the Institute of Behavioral Science in Medicine, Yonsei University College of Medicine. However, restrictions apply to the availability of these data, which were used under license for the current study, and hence not publicly available.

TABLE 2. The architectures of different model subtypes for YUM/ABIDE datasets, each column represents a model whose architecture is arranged from top to bottom in the order of rows

Model Subtypes									
1-1	1-2	2-1	2-2	2-3	2-4	3-1	3-2	4-1	4-2
2DCNN(128/256,3,1)	3DCNN(32,3,1)	N × {2DSTN()}	N × {2DSTN()}	3DSTN()	3DSTN()	3DSTN()*	3DSTN()*	2DCNN(64,3,1)*	3DCNN(32/16,3,1)*
2DMP(2,2)	3DMP(2,2)	2DCNN(16,3,1)	3DCNN(16,3,1)	2DCNN(16,3,1)	3DCNN(16,3,1)	2DCNN(16,3,1)*	3DCNN(32,3,1)*	3DMP(2,2)*	3DMP(2,2)
2DCNN(128/256,3,1)	3DCNN(128,3,1)	2DCNN(16,3,1)	3DMP(2,2)	2DCNN(16,3,1)	3DMP(2.2)	2DCNN(16,3,1)*	3DMP(2,2)*	2DCNN(64,3,1)*	3DCNN(128/64,3,1)
2DMP(2,2)	3DMP(2,2)	FC(1024)	3DCNN(16,3,1)	FC(1024)	3DCNN(16,3,1)	RNN_1(128)	3DCNN(16/64,3,1)*	2DMP(2,2)*	3DMP(2,2)
2DCNN(64/128,3,1)	3DCNN(64,3,1)	FC(2)	3DMP(2,2)	FC(2)	3DMP(2,2)	RNN_2(128)	3DMP(2,2)*	2DCNN(32,3,1)*	3DCNN(32,3,1)
2DMP(2,2)	3DMP(2,2)		FC(1024)		FC(1024)	FC(1024)	RNN_1(128)	2DGAP()*	3DMP(2,2)
DROP(0.2)	DROP(0.2)		FC(2)		FC(2)	FC(2)	RNN_2(128)	FC(1)*	3DGAP()
FC(1024)	FC(1024)						FC(1024)	FC(2)	DROP(0.2)
DROP(0.2)	DROP(0.2)						FC(2)		FC(2)
FC(2)	FC(2)								

TABLE 3. The highlighted regions and MNI coordinates of 2D CAM

Slice	p-value[a]	Original Coordinates	MNI Coordinates	Brain Region
#74	0.41	[134, 170, 74]	[6,25,9]	[Genu of Corpus Callosum Right]
	0.70	[168, 138, 74]	[35,-3,6]	[External Capsule Right]
	0.98	[118, 106, 74]	[-9,-30,2]	[Thalamus Left]
	0.08	[136, 170, 74]	[5,-45,0]	[Cerebellum Right]
#78	0.01	[134, 168, 78]	[6,22,13]	[Lateral Ventricle Frontal Right]
	0.05	[134, 120, 78]	[5,-18,7]	[Thalamus Right]
	0.91	[166, 106, 78]	[32,-31,6]	[Retrolenticular Part of Internal Capsule Right]
	2.78e-04	[120, 104, 78]	[-7,-32,5]	[Thalamus Left]
#104	0.61	[120, 136, 104]	[-7,-8,35]	[Cingulum Left]
	2.76e-05	[88, 104, 104]	[-35,-35,31]	[Superior Longitudinal Fasciculus Left]
	0.02	[136, 104, 104]	[6,-35,31]	[Posterior Cingulate Gyrus Right]

[a] The p-value is calculated by t-test of the values between all the samples in the autism group and the control group at the local maximum point

TABLE 4. Performance comparison of various types of models on YUM and ABIDE datasets

Type	Model Subtype	YUM(μ,σ)	ABIDE(μ,σ)
1-1	3D input+2DCNN	**(0.89,0.04)**	(0.61,0.01)
1-2	3D input+3DCNN	(0.88,0.05)	**(0.64,0.01)**
2-1	2D input+2DCNN+2DSTN	**(0.87,0.03)**	(0.59,0.01)
2-2	2D input+3DCNN+2DSTN	(0.84,0.04)	N/A[b]
2-3	3D input+2DCNN+3DSTN	(0.82,0.04)	(0.57,0.03)
2-4	3D input+3DCNN+3DSTN	(0.85,0.05)	**(0.60,0.01)**
3-1	3D input+2DCNN+3DSTN+RNN	(0.82,0.05)	(0.55,0.01)
3-2	3D input+3DCNN+3DSTN+RNN	**(0.86,0.04)**	**(0.56,0.02)**
4-1	2D input+2DCNN+CAM	(0.84,0.06)	N/A[c]
4-2	3D input+3DCNN+CAM	**(0.86,0.03)**	**(0.56,0.01)**
5-1	3D input+RAM+loc	(0.87,0.03)	(0.58,0.00)
5-2	3D input+RAM+noloc	(0.88,0.01)	(0.58,0.00)
5-3	3D input+RAM+loc+fc	**(0.90,0.03)**	**(0.59,0.00)**
5-4	3D input+RAM+rand	(0.86,0.03)	(0.57,0.00)

[b,c] Learning failed (classification accuracy below 0.5)

TABLE 5. Demographic & Clinical Characteristics of YUM Participants

Phenotypic index	SCQ<15 (n=35)	SCQ≥15 (n=49)	p-value[d]
Male	29 (38.9%)	37 (44.4%)	N/A[f]
Female	6 (61.1%)	12 (55.6%)	N/A[g]
Age	29.4 ± 11.6[d]	30.1 ± 5.3	0.52
SCQ	10.3 ± 3.3	20.12 ± 3.7	<0.01
SMS	65.2 ± 14.3	52.8 ± 13.0	<0.01

[d] Values are mean(μ)± standard deviation(σ)

[e] Chi-square for categorical variable and independent t-test for continuous variable

[f,g] No need to calculate

TABLE 6. Demographic & Clinical Characteristics of ABIDE Participants

Phenotypic index	Autism (n=521)	Control (n=593)
Male	444 (51.9%)	412 (48.1%)
Female	77 (29.8%)	181 (70.2%)
Age	29.4 ± 11.6	30.1 ± 5.3
FIQ	N/A[h]	N/A[i]

[h,i] Due to missing values

V. RESULTS

The crossed-out cells refer to simulation conditions that cannot be run on a standard GPU server due to tremendously high computation costs. 3D and 2D input: both a whole and a single slice of the given MRI image were given as input to the classifier, respectively. CNN: a convolutional neural network, STN: a spatial transformer network, RNN: a recurrent neural network, CAM: a class activation mapping, RAM: a recurrent attention model, loc: a local constraint where an input space was confined to the brain area for the sake of efficiency of learning, noloc: a local constraint was not applied. fc: a fully connected network, rand: random location in each step of RAM. Full details of each model are found in the methods section.

A. TRAINING VARIOUS TYPES OF NEURAL NETWORKS FOR ASD CLASSIFICATION

We ran large-scale simulations to compare the performance of 14 unique versions of models in five different categories. We employed various types of classification tech-

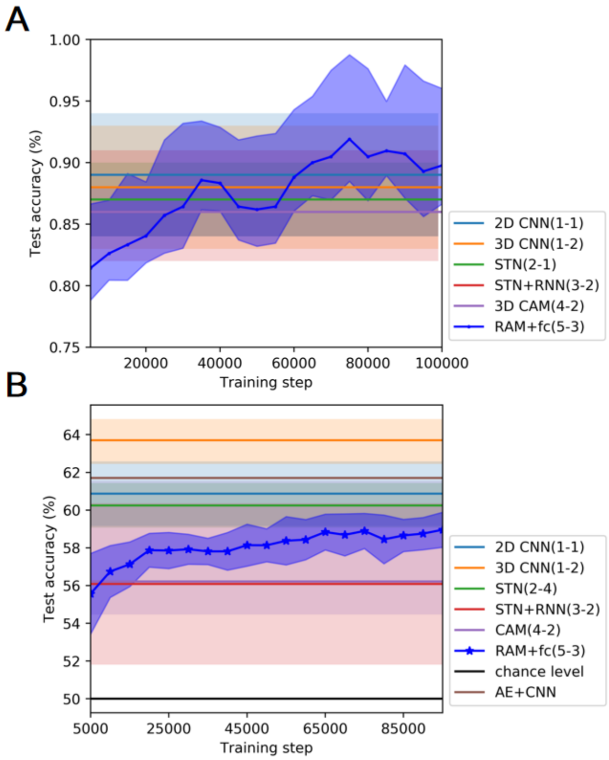


FIGURE 3. Test accuracy of models as a function of training steps. The test accuracy was computed using 10-fold cross validation on the (A) the YUM and (B) the ABIDE, respectively. The names of the models are based on Table 4. For example, RAM+fc refers to the model 5-3 in Table 4. AE+CNN refers to the auto-encoder+CNN model used in [42], and the 50% horizontal line refers to a chance level. The mean test accuracy was recorded every 3000 and 5000 training steps. The shaded area represents a 95% confidence interval.

niques: (A) an invariant method (the Convolution Neural Network (CNN)); (B) a feature learning method (the Spatial Transformer Network (STN)); (C) a feature visualization method (the Classification Activation Mapping(CAM)); (D) a sequence learning model (the Recurrent Neural Network (RNN)); (E) a sequential feature learning model (the Recurrent Attention Model (RAM)); and (6) a generic class of neural networks (the Fully-Connected Network (FC)).

Table 4 shows the details of each combination and the corresponding test accuracy during 10-fold cross-validation (CV). The first four categories are based on an invariant method (CNN) combined with various feature visualization techniques (STN and CAM), whereas the fifth type is based on a sequence learning model (RAM).

The YUM sample consists of 84 subjects (3yr-11yr) with MRI and Social Communication Questionnaire (SCQ) data (see Table 5). Two pediatric psychiatrists at Yonsei University Severance Hospital diagnosed the children as ASD based on DSM-V (see Methods for complete details).

We divided the data into two groups: low and high SCQ, with an SCQ score of 15 set as the threshold (Table 5). The ABIDE dataset is an open-source MRI data repository for autism research (see Methods for more details). The classification accuracy as a function of training epochs is shown in Fig. 3.

We found that the 2D/3D CNN and the RAM performed the best for the YUM dataset, whereas a simple 3D CNN performed the best for the ABIDE dataset (see Fig. 3 and Table 4). Note that the 3D CNN model outperforms the model reported in the previous study [42].

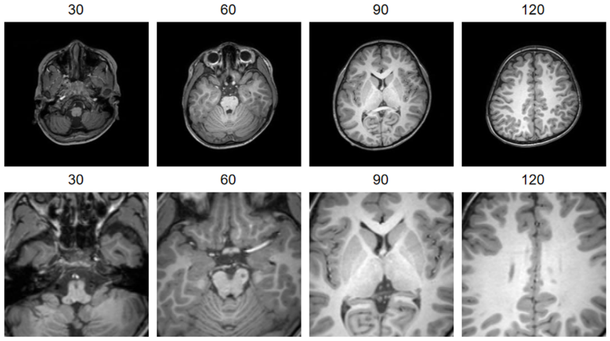


FIGURE 4. Visualization of the features learned by spatial transformer networks. A comparison of slices of the original and the transformed MRI by STN for model type 2 and 3. Images in the first row and the second row are the slices of an original MRI and transformed version, respectively. Note that during training, the STN learned that in order to improve classification accuracy, it would be best to crop the middle cube and resize it to the original size 256×256×256. We selected four representative slices of the MRI. The number on top of each image represents the slice number (z-axis).

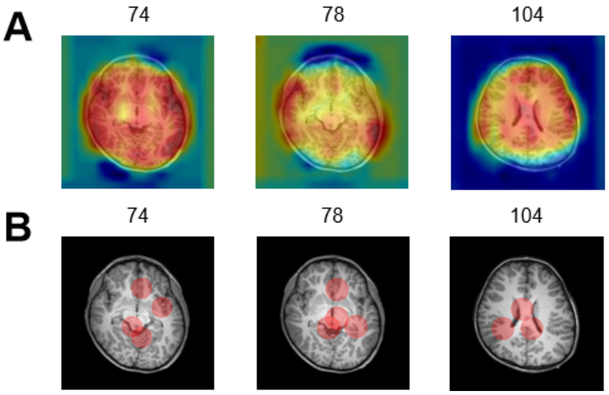


FIGURE 5. Visualization of the feature learned by class activation mapping. The heat maps generated by the CAM and the corresponding local maxima (red dots) for model 4-1 are superimposed on an input brain structure image. Note that to improve computational efficiency and preclude the adverse boundary effect of the model's convolution kernels on CAM results, the results were confined to the region where the brain images are located. To ensure the reliability of the simulation, we acquired the CAM results by running ten cross-validation experiments. For (B), the local maxima are discovered within ten voxels. Refer to Table 3 for the full list of highlighted regions and corresponding MNI coordinates.

B. TRAINING VARIOUS TYPES OF NEURAL NETWORKS FOR ASD CLASSIFICATION

In order to examine which set of input features contributed significantly to the models while categorizing the subjects, we implemented two types of models, each with different characteristics. The first approach was to optimize a linear transformation of input images for classification. We trained the STN on the YUM dataset, a neural network capable of learning an optimal affine transformation of the input image for use in the classification task (refer to model types 2 and 3 in Table 4). The trained STN showed that the optimal input transformation involves cropping the central part of the original 3-dimensional MRI images and then enlarging it to the size of the original image (Fig. 4). This finding suggests that the subcortical structure of the brain might be influential in classification. For the rest of the cases in types 2 and 3, the STN did not learn any meaningful input transformation (data not shown).

The second approach involves training an invariant classifier, such as the convolutional neural networks (CNN) before visualizing the input features that contribute to the model's meaningful classification. We adopted the class activation mapping (CAM) algorithm, which distinguishes a group of informative features from others in the given input. This algorithm estimates the degree of each feature's contribution to the classification (refer to the model type 4 in Table 4). In our work, we have implemented the CAM to create a heat map representing the extent to which the corresponding pixel value contributes to the CNN's classification. We stacked an input image for which the model makes an accurate prediction and its corresponding heat map to visually highlight a particular region of the image that contributes significantly to the model's classification. True positive data are explicitly selected as inputs for the CAM. The heat maps are generated by combining every output of each CAM result for each sample corresponding to the model (4-1) (Fig. 5). Interestingly, local maxima were found in subcortical areas, including the head and the tail of the caudate nucleus (slice #78). A few local maxima also were found in the cortical area, including insular and inferior frontal gyrus (slice #74). Another interesting observation is that the local maxima also includes brain structures with heavy connections, such as claustrum that connects subcortical to cortical areas (slice #74) and corpus callosum that connects the two hemispheres (slice #104). To prevent boundary effect misinterpretations of the model's convolution kernels on CAM results, we excluded the top and bottom eight slides from analysis. Note that most of these brain regions are implicated in decision making, learning, and inhibitory control. One interesting possibility is that these structural differences can contribute to atypical behavior in people with autism spectrum disorders. Note that unlike model (4-1), model (4-2) seems to suffer from an overfitting issue. This issue culminated in less reliable CAM results, which do not warrant discussion. We were not able to apply the CAM to the ABIDE dataset due to impaired visualization of the classification performance, signifying that accuracy did not exceed the chance level.

C. UNDERSTANDING THE STRATEGY BEHIND THE DECISION MAKING

The models that belong to the first four types (types 1 to 4) adhere to single-shot classification, directly predicting the class label for the entire input image. Although the CAM has remarkable ability in visualizing a correlative basis, it lacks the capability to describe causalities between the features of the input image data. In order to discover the optimal strategies to use for accurate classification, we used a recurrent attention model that learns a sequence of voxels (partial brain regions) that the model needs to consider during classification. An optimized sequence can be considered as a set of aptly ordered readouts of brain structures, which ultimately serve as an effective guide for classifying the data. This approach corresponds to the models belonging to type 5.

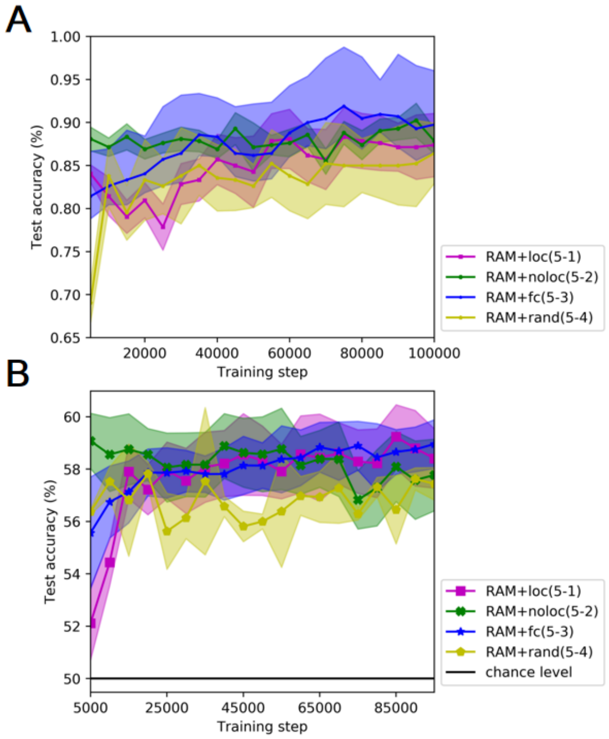


FIGURE 6. Test accuracy for recurrent attention models (RAM). Test accuracy as a function of training epochs for (a) the YUM and (b) the ABIDE. The model RAM+fc (5-3) refers to the model 3D input+RAM+loc+fc (5-3) (Table 4). The test accuracy was measured over 10-fold cross validation. The average test accuracy, indicated by colored dots, was recorded every 3k and 5k training steps for the YUM and the ABIDE, respectively. The training continued until reaching maximum 100k steps. The shaded area represents 95% confidence interval.

All of the type-5 models rapidly identified the optimal input sequences for classification and exceeded 70% accuracy within the first 150K training steps (Fig. 6). For both datasets, a successfully trained model shows a relatively stronger tendency to identify the subcortical structure, including BG (Fig. 7). To formally quantify this effect, we computed the ratio of overlap between the model's attention boxes and basal ganglia (BG) (Fig. 8).

VI. DISCUSSION

We investigated how models comprised of deep neural networks can be applied to identifying individuals with a complex psychiatric disorder such as ASD. The overall

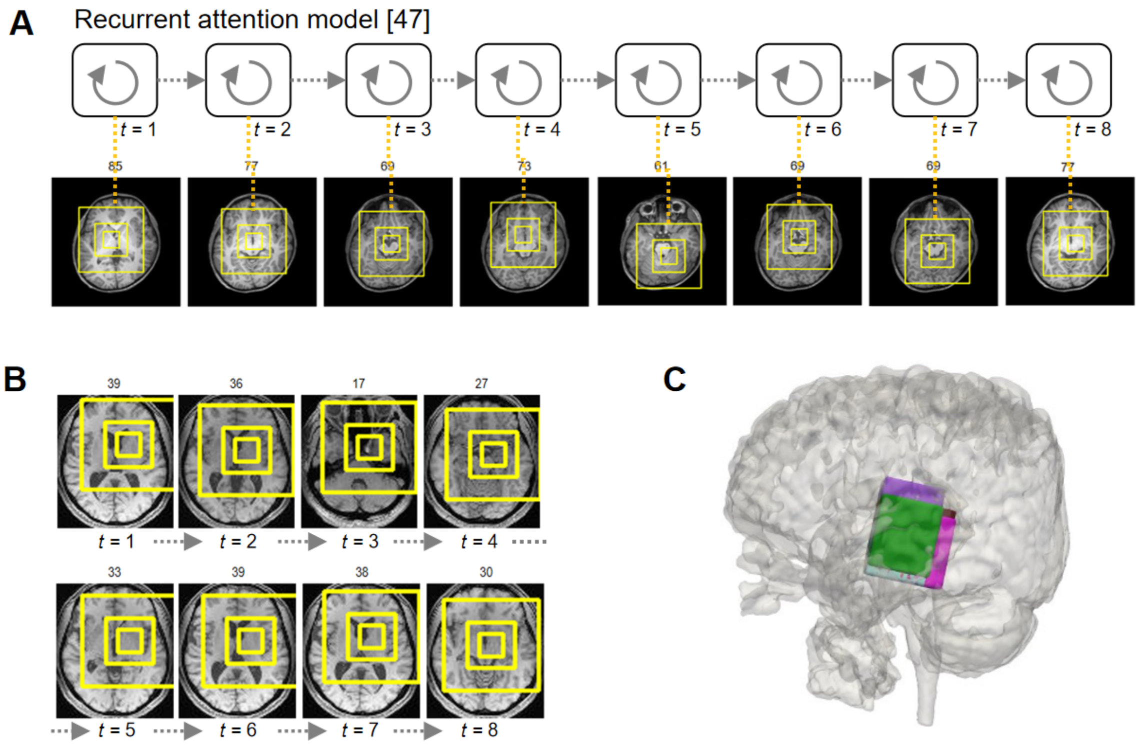


FIGURE 7. Visualization of logics of classification learned by recurrent attention model (RAM). Shown are examples of sequences of brain areas to which the RAM (model subtype: 5-3) attended for classification; (A) the YUM and (B) the ABIDE. The RAM model outputs a classification label after taking each partial 3D image in sequence. The number above each image denotes the slice number.A set of three yellow rectangles indicates a middle slice of the 3D attention box (cubic). (C) Three dimensional view of the attention boxes (A), color-coded by the sequence indices.

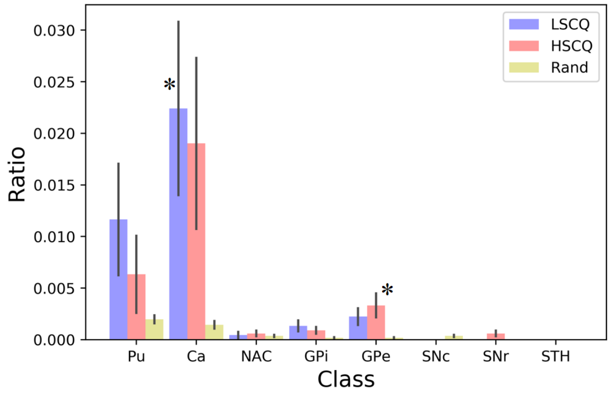


FIGURE 8. Importance of the subcortical structure for ASD classification. Shown is the ratio of overlaps between attention boxes of the recurrent attention model and eight regions of subcortical nuclei, including Putamen (Pu), Caudate (Ca), Nucleus Acumbens (NAC), Globus Pallidus internal (GPi), Globus Pallidus external (GPe), Substantia Nigra compacta (SNc), Substantia Nigra reticulata (SNr), Subthalamic Nucleus(STH). We used a binarized mask extracted from a probabilistic subcortical nuclei mask with the threshold probability 0.5 [58]. The information of attention boxes was extracted from the recurrent attention model trained on the YUM. The blue and red bar refers to the low (LSCQ) and the high SCQ group (HSCQ), respectively. The yellow bar refers to the case with random sampling. The error bar represents 95% confidence interval. The asterisk indicates statistical significance (p<0.05; paired t-test between LSCQ/HSCQ and Rand).

architecture is summarized in Fig. 1. We primarily used the CNN and RNN as analysis and diagnosis tools, building them with various architectures. We measured the performance of every model on classification tasks, with each task using a different MRI dataset.

Unlike conventional approaches that extract morphological features using traditional algorithms, we directly fitted neural networks to the original MRI voxel data, finding the structural difference between the autism and control groups. Our end-to-end training regime does not require extraction of human morphological feature information, reducing the risk of missing information and causing errors in the extraction process.

Note that this paper aims not only to reliably enhance classification accuracy, but also and more importantly, to explore structural and strategic ASD evidence. We achieve this goal by using a relatively large sample size and by exploring a variety of different model versions, including 2D/3D CNN, STN, and RAM. For example, RAM provides the logic of classification (Fig. 6); however, the ABIDE dataset's test accuracy is slightly lower than the best version. There are several reasons why it is challenging for YUM and ABIDE to achieve consistent accuracy:

- Data variability: ABIDE is a collection of data from more than 20 institutions, each with different scanners, scanning protocols, and configuration parameters, making image features very different from those included in the YUM data. Transferring ABIDE data to the MNI152 standard template unavoidably caused image variability.

On the other hand, the YUM data set had relatively smaller variability because it was collected by the same facility. This fact might explain why the RAM showed strong performance for the YUM in comparison to the versions based on the invariant method, such as 2D or 3D CNN.

- Sample size: The sample size of ABIDE involves more than 1000 images, whereas the YUM contains only 84. It is generally known that CNN models show reliable performance when the sample size is sufficiently large (ABIDE). However, attention-based models, such as RAM, hold an advantage when the sample size is very small (YUM).
- Structural heterogeneity: ABIDE includes a very broad age range for patients with autism, implying substantially higher heterogeneity than YUM (refer to both Table 5 and Table 6).
- Spatial resolution: The YUM consists of high resolution sMRI. The spatial resolution of YUM is higher than that of ABIDE.
- Class labeling: The method for labeling the ABIDE data differs slightly from that of the YUM data, which relies on the SCQ (Social Communication Questionnaire) index.

We built the CAM and a diagnosis sequence generator on top of the CNNs and the RNNs, respectively. The CAM numerates the contribution degree of each input. In other words, the algorithm computes a value that represents how often and how strongly the model refers to a particular feature during the classification tasks. Psychiatric physicians can use this type of analysis tool to identify significant brain regions during the diagnosis process. We also have run both the grad-CAM and the guided grad-CAM on our dataset. Despite much effort to fine-tune these models, visualization results are slightly noisier and less reliable than those done with CAM. The input of 3D CAM and 2D CAM differ due to differences in structure, 3D volume and 2D slice, respectively. This variance also explains why 3D and 2D CAM offer different results in some areas. That being said, based on the overall statistical analyses, we found that the results from these two models consistently overlapped in the thalamus, caudate nucleus, claustrum, and other subcortical tissue areas. Further, applying the RNN generates an optimized sequence of the brain regions, which can serve as a remarkable index for clinicians. The generator provides rigorously ordered brain regions to aid in diagnosis. Such structural and strategic clinical models may be state-of-the-art indicators of ASD. Using these models in clinical settings may positively impact individual patients while increasing efficiency and economic benefits for the community at large.

The major regions in the classification were subcortical structures, including the BG. The BG, which itself consists of the striatum, caudate nucleus, globus pallidus, and putamen, is a group of subcortical structures involved in motor function as well as learning and memory. BG is suspected to contribute to repetitive and stereotyped behaviors, which is a core symptom domain of autism spectrum disorder. Despite the limited implications of the BG's role in autistic symptoms, there is little evidence from previous high-resolution MRI ($\geq$3T) studies. Our results (Fig. 8) strongly support the idea that the BG area could be a potential biomarker of autism.

To the best of our knowledge, Ghiassian and Sen's papers are the only two demonstrations using automated learning methods to classify the autism patient using extensive databases. There are a few differences between our model and the models used in previous studies. Firstly, Ghiassian's study relies on a hand-crafted histogram of oriented gradients, which may be prone to subjective bias. In contrast, we employed an end-to-end training regime for classification. Secondly, unlike Sen's study, our study adopted autoencoders for data reconstruction. We were able to avoid weights transfer, which usually is used in the classification task. Thus, the filter number does not necessarily match the number of units in the hidden layer of the sparse autoencoder. Third, we used a 3D-CNN that learns the complex spatial patterns of features. This setting reflects our perspective on a volume or thickness of gray and white matter such that they can be good indicators of ASD. Note that our model outperforms the 2D-CNN by 2.8% in overall accuracy.

The reported classification accuracy may be considered inadequate to reach the level for clinical utility. Despite this technical insufficiency, our study provides a useful protocol for visualizing elements with neural networks learning from the data, as well as perceiving their relationships. These findings will allow profound clinical insights into ASD diagnosis. Our study blazes a trail in discovering structural and strategic evidence for acknowledging complex psychiatric symptoms, thereby guiding clinicians in refining currently-available diagnostic tools.

AUTHOR CONTRIBUTIONS

F.K. Ke developed the study concept and implemented the paradigm under supervision of S.W. Lee. S.J. Choi collected the data under supervision of K.A. Cheon. F.K. Ke and Y.H. Kang devised analyses methods for the study. F.K. Ke, Y.H. Kang and S.W. Lee analyzed the data, F.K. Ke and Y.H. Kang drafted the manuscript, and S.W. Lee and K.A. Cheon commented and revised it. All authors approved the final version of the manuscript for submission. The source code in this paper can be found in https://github.com/brain-machine-intelligence/Autism-Classification.

ADDITIONAL INFORMATION

Competing Interests: the authors declare no competing interests.

REFERENCES

[1] K. Khowaja, B. Banire, D. Al-Thani, M. T. Sqalli, and S. S. Salim, "Augmented reality for learning of children and adolescents with autism spectrum disorder (asd): A systematic review," IEEE Access, vol. 8, pp. 78779 – 78807, 2020.

[2] B. Jon, W. Lisa, C. D. L., M. M. J., D. Julie, W. Zachary, K.-S. Margaret, Z. Walter, and R. Cordelia, "Prevalence of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2014," Mmwr Surveillance Summaries, vol. 67, no. 6, pp. 1–23, 2018.

[3] K. Y. Shin, L. B. L., K. Y. Joo, F. Eric, L. Eugene, L. E. Chung, C. K. Ah, K. S. Jeong, K. Y. Key, and L. H. Kyung, "Prevalence of autism spectrum disorders in a total population sample," Am J Psychiatry, vol. 168, no. 9, pp. 904–912, 2011.

[4] S. Mostafa, L. Tang, and F.-X. Wu, "Diagnosis of autism spectrum disorder based on eigenvalues of brain networks," IEEE Access, vol. 7, pp. 128474 – 128486, 2019.

[5] E. C., B. S.Y., and M. D.G., "Neuroimaging in autism spectrum disorder: brain structure and function across the lifespan," Lancet Neurol, vol. 14, no. 11, pp. 1121–1134, 2015.

[6] C. Wang, Z. Xiao, B. Wang, and J. Wu, "Identification of autism based on svm-rfe and stacked sparse auto-encoder," IEEE Access, vol. 7, pp. 118030 – 118036, 2019.

[7] G. Fan, Y. Chen, Y. Chen, M. Yang, and T. Liu, "Abnormal brain regions in two-group cross-location dynamics model of autism," IEEE Access, vol. 8, pp. 94526 – 94534, 2020.

[8] T. Akter, Satu, Khan, M. H. Ali, and M. A. Moni, "Machine learning-based models for early stage detection of autism spectrum disorders," IEEE Access, vol. 7, pp. 166509 – 166527, 2019.

[9] Arbabshirani, M. R., P. Sergey, S. Jing, and C. V. D., "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," Neuroimage, vol. 145, no. Pt B, p. 137, 2016.

[10] C. V.D., A. T., G. N.R., P. J.J., and P. G.D., "Method for multimodal analysis of independent source differences in schizophrenia: Combining gray matter structural and auditory oddball functional data," Human Brain Mapping, vol. 27, no. 1, pp. 47–62, 2006.

[11] W. T., B. J. K., B. C. F., F. B., and M. A. F., "From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics," Neurosci Biobehav Rev, vol. 57, pp. 328–349, 2015.

[12] P. P. D. M. Oliveira, R. Nitrini, G. Busatto, C. Buchpiguel, and E. Amaro, "Use of svm methods with surface-based cortical and volumetric subcortical measurements to detect alzheimer's disease," Journal of Alzheimers Disease Jad, vol. 19, no. 4, pp. 1263–1272, 2010.

[13] S. T. Yang, L. Jiann-Der, T. C. Chang, C. H. Huang, J. J. Wang, H. Wen-Chuin, H. L. Chan, W. Yau-Yau, and K. Y. Li, "Discrimination between alzheimer's disease and mild cognitive impairment using som and pso-svm," Computational and Mathematical Methods in Medicine, vol. 2013, pp. 1–10, 2013.

[14] P. Claudia, S. J. Teipel, A. Oswald, C. B., T. Meindl, J. Mourao-Miranda, A. W. Bokde, H. Hampel, and M. Ewers, "Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease," Neuroimage, vol. 50, no. 1, pp. 162–174, 2009.

[15] M. R. Sabuncu and K. V. Leemput, "The relevance voxel machine (rvoxm): a self-tuning bayesian model for informative image-based prediction," IEEE Trans Med Imaging, vol. 31, no. 12, pp. 2290–2306, 2012.

[16] S. P. Koch, C. Hägele, J.-D. Haynes, A. Heinz, F. Schlagenhauf, and P. Sterzer, "Diagnostic classification of schizophrenia patients on the basis of regional reward-related fmri signal patterns," PLoS One, vol. 10, no. 3, p. e0119089, 2015.

[17] M. Bleich-Cohen, S. Jamshy, H. Sharon, R. Weizman, N. Intrator, M. Poyurovsky, and T. Hendler, "Machine learning fmri classifier delineates subgroups of schizophrenia patients," Schizophr Res., vol. 160, no. 1-3, pp. 196–200, 2014.

[18] M. Y. Park and T. Hastie, "L1-regularization path algorithm for generalized linear models," Journal of the Royal Statistical Society, vol. 69, no. 4, pp. 659–677, 2007.

[19] J. B. A., M. Benson, M. Keith, C. David, K. Kerstin, and S. J. Douglas, "Brainstem abnormalities in attention deficit hyperactivity disorder support high accuracy individual diagnostic classification," Human Brain Mapping, vol. 35, 2014.

[20] Y. Shimizu, J. Yoshimoto, S. Toki, M. Takamura, S. Yoshimura, Y. Okamoto, S. Yamawaki, and K. Doya, "Toward probabilistic diagnosis and understanding of depression based on functional mri data analysis with logistic group lasso," PLoS One, vol. 10, p. e0123524, 2015.

[21] P. M. J., A. Carmen, P. J. C., E. K. L., R. C. F., and A. H. J., "Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction," International Journal of Geriatric Psychiatry, vol. 30, pp. 1056–1067, 2015.

[22] H. G., D. L., Y. D., and et al, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," Signal Processing Magazine IEEE, vol. 29, no. 6, pp. 82–97, 2012.

[23] T. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8614–8618, 2013.

[24] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation," Computer Science, 2014.

[25] S. Ilya, V. Oriol, and L. Q. V., "Sequence to sequence learning with neural networks," Advances in neural information processing systems, 2014.

[26] K. Alex, S. I., and H. G., "Imagenet classification with deep convolutional neural networks," in NIPS, vol. 25, pp. 1097–1105, 2012.

[27] J. Tompson, A. Jain, Y. Lecun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," Eprint Arxiv, vol. 25, pp. 1799–1807, 2014.

[28] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," Annu Rev Biomed Eng., vol. 19, pp. 221–248, 2017.

[29] S. Liu, S. Liu, W. Cai, S. Pujol, R.Kikinis, and D. Feng, "Early diagnosis of alzheimer's disease with deep learning," in 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp. 1015–1018, 2014.

[30] H. I. Suk, S. W. Lee, and D. Shen, "Latent feature representation with stacked auto-encoder for ad/mci diagnosis," Brain Structure and Function, vol. 220, no. 2, pp. 841–859, 2013.

[31] J. Kim, V. D. Calhoun, E. Shim, and J. H. Lee, "Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia," Neuroimage, vol. 124, pp. 127–146, 2015.

[32] N. J. Minshew and J. B. Payton, "New perspectives in autism. part 2: the differential diagnosis and neurobiology of autism," Current Problems in Pediatrics, vol. 18, no. 11, pp. 618–694, 1988.

[33] L. Wing, "The autistic spectrum," Lancet, vol. 350, no. 9093, pp. 1761–6, 1997.

[34] N. Akshoomoff, C. Lord, A. J. Lincoln, R. Y. Courchesne, R. A. Carper, J. Townsend, and E. Courchesne, "Outcome classification of preschool children with autism spectrum disorders using mri brain measures," Journal of the American Academy of Child and Adolescent Psychiatry, vol. 43, no. 3, pp. 349–357, 2004.

[35] Y. Jiao, R. Chen, X. Ke, K. Chu, Z. Lu, and E. H. Herskovits, "Predictive models of autism spectrum disorder based on brain regional cortical thickness," Neuroimage, vol. 50, no. 2, pp. 589–599, 2010.

[36] C. Ecker, V. Rocha-Rego, P. Johnston, J. Mourao-Miranda, A. Marquand, E. M. Daly, M. J. Brammer, C. Murphy, and D. G. Murphy, "Investigating the predictive value of whole-brain structural mr scans in autism: A pattern classification approach," Neuroimage, vol. 49, no. 1, pp. 44–56, 2010.

[37] S. Ferm?-N, H. Rosemary, S. Michael, J. M. G?3Rriz, R. R. Javier, C. G. Puntonet, P. Christophe, C. Lindsay, B. C. Simon, and S. John, "Identifying endophenotypes of autism: A multivariate approach," Frontiers in Computational Neuroscience, vol. 8, no. 60, p. 60, 2014.

[38] C. Ecker, A., Marquand, J., Mourao-Miranda, P., Johnston, E., and M. and, "Describing the brain in autism in five dimensions–magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach," Journal of Neuroscience, vol. 30, pp. 10612–10623, 2010.

[39] I. Gori, A. Giuliano, F. Muratori, I. Saviozzi, P. Oliva, R. Tancredi, A. Cosenza, M. Tosetti, S. Calderoni, and A. Retico, "Gray matter alterations in young children with autism spectrum disorders: Comparing morphometry at the voxel and regional level," Journal of Neuroimaging Official Journal of the American Society of Neuroimaging, vol. 25, no. 6, pp. 866–874, 2015.

[40] O. Demirci, V. P. Clark, V. A. Magnotta, N. C. Andreasen, J. Lauriello, K. A. Kiehl, G. D. Pearlson, and V. D. Calhoun, "A review of challenges in the use of fmri for disease classification / characterization and a projection pursuit application from a multi-site fmri schizophrenia study," Brain Imaging and Behavior, vol. 2, no. 3, pp. 207–226, 2008.

[41] I. Nouretdinov, S. G. Costafreda, A. Gammerman, A. Chervonenkis, V. Vovk, V. Vapnik, and C. H. Y. Fu, "Machine learning classification with confidence: Application of transductive conformal predictors to mri-based diagnostic and prognostic markers in depression," Neuroimage, vol. 56, no. 2, pp. 809–813, 2011.

[42] G. Sina, G. Russell, P. Jin, M. R. G. Brown, and H. Leontios, "Using functional or structural magnetic resonance images and personal characteristic data to identify adhd and autism," vol. 11, no. 12, p. e0166934, 2016.

[43] S. Bhaskar, N. C. Borle, G. Russell, M. R. G. Brown, and B. C. Bernhardt, “A general prediction model for the detection of adhd and autism using structural and functional mri,” Plos One, vol. 13, no. 4, pp. e0194856–, 2018.

[44] M. Leming and J. Suckling, “Deep learning on brain images in autism: What do large samples reveal of its complexity?,” in International Work-Conference on the Interplay Between Natural and Artificial Computation(IWINAC), pp. 389–402, 2019.

[45] P. W.H.L., M. A., and J. R. Sato, “Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study,” Human Brain Mapping, vol. 40, 2019.

[46] R. C. Craddock, G. James, P. E. H. III, X. P. Hu, and H. S. Mayberg, “A whole brain fmri atlas generated via spatially constrained spectral clustering,” Human Brain Mapping, vol. 33, no. 8, p. 1914–1928, 2012.

[47] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, “Identification of autism spectrum disorder using deep learning and the abide dataset,” Neuroimage Clinical, p. S2213158217302073, 2017.

[48] S. Alexander, R. Kong, E. M. Gordon, T. O. Laumann, Z. Xi-Nian, A. J. Holmes, S. B. Eickhoff, and Y. B. T. Thomas, “Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri,” Cerebral Cortex, no. 9, p. 9, 2017.

[49] A. E. Gazzar, L. Cerliani, G. van Wingen, and R. M. Thomas, “Simple 1-d convolutional networks for resting-state fmri based classification in autism,” in 2019 International Joint Conference on Neural Networks (IJCNN), 2019.

[50] X. Yang, M. Islam, and A. Khaled, “Functional connectivity magnetic resonance imaging classification of autism spectrum disorder using the multisite abide dataset,” in 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), pp. 1–4, 2019.

[51] A. D. Martino, C.-G. Yan, L. Q, D. E, C. F. X, A. K, and et al, “The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism.,” Molecular Psychiatry, vol. 19, pp. 659–667, 2014.

[52] J. Ashburner, G. Barnes, C. Chen, J. Daunizeau, G. Flandin, K. Friston, S. Kiebel, J. Kilner, V. Litvak, and R. Moran, “Spm8 manual the fil methods group (and honorary members),” 2013.

[53] V. S. Fonov, A. C. Evans, R. C. Mckinstry, C. R. Almli, and D. L. Collins, “Unbiased nonlinear average age-appropriate brain templates from birth to adulthood,” Neuroimage, vol. 47, pp. S102–S102, 2009.

[54] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” Adv Neural Inf Process Syst, vol. 28, pp. 2017–2025, 2015.

[55] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, “Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation,” Medical Image Analysis, vol. 36, p. 61, 2016.

[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929, 2015.

[57] Mnih, Volodymyr, Heess, Nicolas, A. Graves, and k. kavukcuoglu, “Recurrent models of visual attention,” in Advances in Neural Information Processing Systems 27 (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 2204–2212, Curran Associates, Inc., 2014.

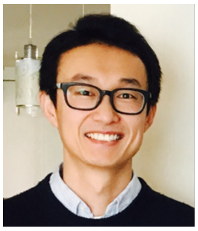
[58] W. M. Pauli, A. N. Nili, and J. M. Tyszka, “A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei,” Scientific Data, vol. 5, p. 180063, 2018.

FENGKAI KE In 2016, Fengkai Ke was graduated from the school of mechanical engineering and science, Huazhong University of Science and Technology (HUST) with a doctor’s degree. In 2018, he worked as a postdoctoral researcher at Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. Dr. Fengkai Ke is now working in the school of mechanical engineering, Hubei University of Technology, Wuhan, Hubei Province, China. During his doctoral and postdoctoral studies, he participated in a number of national and provincial natural science fund projects, and has published more than ten authoritative journals and international conference papers. His main research fields are in deep learning, reinforcement learning, medical image processing, etc.

SEUNG JIN CHOI Seung Jin Choi is affiliated with the Department of Child and Adolescent Psychiatry, Severance Children’s Hospital. Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea. His research interests include neuroimaging studies of Autism Spectrum Disorders and ADHD.

YOUNG HO KANG Young Ho Kang is a Ph.D candidate of Program of Brain and Cognitive Engineering at KAIST. He received his B.Eng (Hons) in Computer Systems Engineering from Department of Computer Science, the University of Manchester in 2017. His research interests include brain-inspired artificial intelligence, computational cognitive neuroscience and computational psychiatry.

KEUN-AH CHEON Dr. Keun-Ah Cheon is a professor in the department of Psychiatry at Yonsei University College of Medicine, Severance Hospital and a Director in Department of Child and Adolescent Psychiatry. She is a core faculty member of Institute of Behavioral Science in Medicine, Yonsei University College of Medicine. She is also a founding director of the Yonsei Autism Laboratory in S. Korea. She received her Ph.D. in Medical Science from Yonsei University Graduate School in 2003. She worked as a visiting scholar at the Developmental Neuroimaging Laboratory in Center for Autism Research in Children’s Hospital of Philadelphia, University of Pennsylvania School of Medicine, USA. She was the recipient of the Donald J Cohen Fellowship Award of the International Association of Child and Adolescent Psychiatry and Allied Professionals. Her research interests include neuroimaging studies of Autism Spectrum Disorders and ADHD.



SANG WAN LEE Sang Wan Lee is an associate professor in the department of Bio and Brain Engineering at KAIST, a core faculty member of Program of Brain and Cognitive Engineering, KAIST Institute for Health, Science, and Technology (KI HST), and KAIST Institute for Artificial Intelligence (KI AI). He is also a founding director of KAIST Center for Neuroscience-inspired AI. He received his Ph.D. in Electrical Engineering and Computer Science from KAIST in 2009. He was a postdoctoral associate at MIT, followed by a Della Martin postdoctoral scholar at Caltech. He was the recipient of the Della-Martin Fellowship, Google Faculty Research Award, and KAIST Songam Distinguished Research Award in 2014, 2017, and 2019, respectively. His research interests include neuroscience-inspired artificial intelligence and decision neuroscience.