Hindawi
Complexity
Volume 2020, Article ID 1357853, 9 pages
https://doi.org/10.1155/2020/1357853

Research Article

Multichannel Deep Attention Neural Networks for the Classification of Autism Spectrum Disorder Using Neuroimaging and Personal Characteristic Data

Ke Niu,[1,2] **Jiayang Guo**,[3] **Yijie Pan**,[4] **Xin Gao**,[5] **Xueping Peng**,[2] **Ning Li**,[1] **and Hailong Li**[6]

[1]*Computer School, Beijing Information Science and Technology University, Beijing 100101, China*
[2]*CAI, School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, Australia*
[3]*Department of Electrical Engineering and Computer Science, University of Cincinnati, Cincinnati, OH 45221, USA*
[4]*Ningbo Institute of Information Technology Application, CAS, Beijing, China*
[5]*Computational Bioscience Research Center (CBRC), Computer Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia*
[6]*Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA*

Correspondence should be addressed to Jiayang Guo; guojy@mail.uc.edu, Xueping Peng; xueping.peng@uts.edu.au, and Hailong Li; hailong.li@cchmc.org

Received 12 June 2019; Revised 1 January 2020; Accepted 4 January 2020; Published 31 January 2020

Guest Editor: Gonzalo Farias

Autism spectrum disorder (ASD) is a developmental disorder that impacts more than 1.6% of children aged 8 across the United States. It is characterized by impairments in social interaction and communication, as well as by a restricted repertoire of activity and interests. The current standardized clinical diagnosis of ASD remains to be a subjective diagnosis, mainly relying on behavior-based tests. However, the diagnostic process for ASD is not only time consuming, but also costly, causing a tremendous financial burden for patients' families. Therefore, automated diagnosis approaches have been an attractive solution for earlier identification of ASD. In this work, we set to develop a deep learning model for automated diagnosis of ASD. Specifically, a multichannel deep attention neural network (DANN) was proposed by integrating multiple layers of neural networks, attention mechanism, and feature fusion to capture the interrelationships in multimodality data. We evaluated the proposed multichannel DANN model on the Autism Brain Imaging Data Exchange (ABIDE) repository with 809 subjects (408 ASD patients and 401 typical development controls). Our model achieved a state-of-the-art accuracy of 0.732 on ASD classification by integrating three scales of brain functional connectomes and personal characteristic data, outperforming multiple peer machine learning models in a k-fold cross validation experiment. Additional k-fold and leave-one-site-out cross validation were conducted to test the generalizability and robustness of the proposed multichannel DANN model. The results show promise for deep learning models to aid the future automated clinical diagnosis of ASD.

1. Introduction

Autism spectrum disorder (ASD) has been estimated to occur in more than 1.6% of children aged 8 across the United States [1]. As a chronic neurological condition, ASD is characterized by impairments in social interaction and communication, as well as by a restricted repertoire of activity and interests [2–5]. Patients with ASD exhibit different levels of impairments, ranging from above average to intellectual disability. In neuroscience, ASD remains a formidable challenge, due to their high prevalence, complexity, and substantial heterogeneity, which require multidisciplinary efforts [6–8]. Although clinical therapies have been developed to treat the symptoms, the diagnosis of ASD

remains to be a challenging task. Currently, behavior-based test is the standard clinical method for diagnosing ASD [9]. However, the diagnostic process for ASD is not only time consuming but also costly [10]. This results in a tremendous financial burden for patients' families. Meanwhile, with this lifetime ASD, the patients may have difficulties in normal socialization and working environments, increasing the overall social costs. Therefore, an automated diagnosis approach is desirable for earlier identification of ASD.

Machine learning is a promising tool for investigating the replicability of patterns across larger, more heterogeneous datasets [11–13]. For automated diagnosis of ASD, personal characteristic (PC) data, such as intelligence quotient (IQ) and Social Responsiveness Scale (SRS) score have been adopted in several studies [14–16]. In the study of ASD, IQ is a type of standard score that is derived from several standardized tests designed to assess human intelligence, and the SRS score includes a 65-item standardized questionnaire regarding behaviors that are associated with ASD [17]. ASD is highly associated with intellectual disability which is mainly measured by IQ. Meanwhile, some studies [18, 19] indicate that IQ discrepancy marks a meaningful phenotype in ASDs. In this way, IQ becomes an important biomarker to classify the ASD.

Neuroimaging data have also been investigated to explore ASD biomarkers in recent decades. To facilitate the ASD research community, Autism Brain Imaging Data Exchange (ABIDE), an international collaborative project, has collected data from over 1,000 subjects (e.g., structure MRI (sMRI), resting-state functional MRI (rs-fMRI), and PC data) and made the whole database publicly available. This provided a common platform to test hypotheses, search key biomarkers, and develop advanced statistical and machine learning algorithms. For example, Ghiassian et al. [20] proposed an automated classifier by combining the histogram of orientated gradients approach for feature extraction from sMRI and rs-fMRI data and support vector machines (SVMs) for decision making. Their method was tested on the ABIDE dataset and achieved 65.0% accuracy on hold-out set. Of late, Sen et al. [21] developed a LEFMS learner, which applies sparse autoencoder to extract features from sMRI and spatial nonstationary independent components on rs-fMRI data. SVM was the utilized to classify ASD and improved accuracy by 0.042. Katuwal et al. [22] applied a random forest classifier to classify ASD and achieved an AUC of 0.61. Adding verbal IQ and age to morphometric features, AUC was improved to 0.68. By introducing hypergraph learning technique, Zu et al. [23] proposed a novel learning method to discover complex connectivity biomarkers that are beyond the widely used region-to-region connections in the conventional brain network analysis.

Deep learning has had a profound impact on many data analytic applications, such as speech recognition, image classification, computer vision, and natural language processing [24]. Based on data-driven feature construction, deep learning provides a new direction for data analytic modelling. Over the past few years, an increasing body of the literature confirmed the success of feature construction using deep learning methods. Deep learning has been demonstrated to outperform traditional machine learning algorithms on numerous recognition and classification tasks [24–29], which inspires the researchers in the ASD community to apply deep learning approaches on ASD classification. Earlier, deep neural networks (DNNs) have been applied to identify ASD patients using rs-fMRI [26]. Their model achieved 70% on accuracy by using the functional connectivity (FC) matrix as features for model training.

Kong et al. [27] constructed individual functional brain networks using the rs-fMRI data from 182 subjects of NYU Langone Medical Center, a data site within ABIDE repository. FC features were used to represent the networks of all subjects and further ranked using F-score. Then, a stacked sparse autoencoder-based DNN model was developed. Significant performance improvement was achieved by comparing the proposed method with two existing algorithms.

More recently, an ASD-DiagNet, a joint learning procedure using an autoencoder and a single layer perceptron, was presented [28]. A data augmentation strategy was also designed for the FC features of functional brain networks based on linear interpolation of available feature vectors to ensure the robust training of the ASD-DiagNet. By evaluating the model on 1035 subjects from 17 different sites of ABIDE repository, ASD-DiagNet achieves 70.1% on the accuracy, 67.8% on sensitivity, and 72.8% on specificity in 10-fold cross validation. In the mode evaluation of individual data centers, ASD-DiagNet outperformed other state-of-the-art methods and increased the accuracy performance up to 20% with a maximum accuracy of 80%.

In this work, we aim to develop a novel deep learning model for automated diagnosis of ASD. Specifically, we proposed a multichannel deep attention neural network, called DANN, by integrating multiple layers of neural networks, attention mechanism, and feature fusion to capture the interrelationships in multimodality data (functional neuroimaging data and PC data) to distinguish ASD patients from typical development controls (TDCs). The attention mechanism-based learning is a type of deep learning which is a recent trend for understanding what part of historical information weighs more in predicting diseases [30, 31]. Taking advantage of large heterogeneous dataset from ABIDE, multiscale brain functional connectomes and PC data were obtained as the features. We systematically evaluated the diagnosis power of our multichannel DANN on ASD classification and compared the performance of the proposed model with peer machine learning models.

The rest of paper is organized as follows. Section 2 describes ASD data and multichannel deep attention neural network. The experimental setup is shown in Section 3, followed by the experimental results and discussion in Section 4. Finally, the conclusion of this work is described in Section 5.

2. Materials and Methods

2.1. Subjects. We collected preprocessed rs-fMRI and PC data from 809 subjects from publicly accessible ABIDE repository, including 408 ASD subjects and 401 TDC subjects. Detailed demographic information of subjects is listed in Table 1. The incidence of ASD between male and female

Table 1: Demographic information of 809 subjects from ABIDE.

Type	Number	Gender (M/F)	Age	FIQ	PIQ	VIQ
ASD	408	330/78	16.47 ± 6.70	110.63 ± 12.67	107.85 ± 13.41	111.17 ± 13.31
TDC	401	352/49	16.80 ± 7.80	105.28 ± 16.64	105.10 ± 17.10	104.60 ± 17.81
p value	—	0.017	0.785	<0.001	0.003	<0.001

ASD: autism spectrum disorder; TDC: typical development control; M: male; F: female; VIQ: the verbal IQ; PIQ: the performance IQ; FIQ: the full-scale IQ. The values are denoted as mean and standard deviation.

subjects is significantly different, and thus the majority of the subjects in ABIDE dataset are male. There is no significant difference between the age of ASD and TDC groups. All three IQ scores had significant difference between two groups. Later, the variables' gender, age, and three IQs were used as PC data in our ASD classification experiments.

2.2. Data Preprocessing. Each of rs-fMRI data has been preprocessed using Configurable Pipeline for the Analysis of Connectomes (CPAC) preprocessing pipeline, which includes slice timing correction, motion realignment, and intensity normalization. Nuisance variable regression was implemented through bandpass filtering and global signal regression strategies to clean confounding variations introduced by heartbeats and respiration, head motion, and low-frequency scanner drifts. Furthermore, boundary-based rigid body and FMRIB's linear and nonlinear image registration tools were used to register functional to anatomical images. Then, both functional and anatomical images were normalized to template space (MNI 152). Three scales of brain functional connectomes were extracted in this work. Mean blood oxygen-level dependent (BOLD) time-series signals for three sets of regions of interests (ROIs), i.e., atlases, including the Automated Anatomical Labeling (AAL) atlas, Harvard-Oxford (HO) atlas, and Craddock 200 (CC200), were calculated. The weights of functional brain connectivity were defined using Pearson's correlation coefficient between any pair of two ROIs. For AAL atlas, each subject was represented by a 90×90 FC adjacency matrix, symmetric along diagonal, in which each entry represents the brain connectivity between each pair of ROIs. Similarly, each rs-fMRI data was also represented by 110×110 and 200×200 symmetric FC adjacency matrices using HO and CC200 atlases, respectively. In addition, from 809 subjects, we obtained five PC data, including sex, handedness, full-scale IQ (FIQ), verbal IQ (VIQ), and performance IQ (PIQ).

2.3. Multichannel Deep Attention Neural Network

2.3.1. Overview Structure. An overview of multichannel DANN is given in Figure 1. It consists of blocks of multichannel inputs, multilayer perceptron (MLP), self-attention, fusion, and aggregation. The various components are described in the following sections.

2.3.2. MLP. The MLP block is composed of 5 layers, which are one dropout layer and four dense layers. The details of the block are shown in Figure 2.

A dropout layer, which prevents overfitting during training the model, is applied on input data, e.g. AAL FC (input size is 4005). The white circle in Figure 2 denotes dropped units according to dropout probability. The dropout layer is followed by four dense layers, whose hidden units are 1024, 512, 128, and 32, respectively, and corresponding activation functions are "elu," "tanh," "tanh," and "relu," respectively.

2.3.3. Self-Attention. The attention is proposed to compute an alignment score between elements from two sources [32]. In particular, given an input FC adjacency matrix, which can be transformed into a FC adjacency sequence, $\mathbf{x} = [x_1, x_2, \ldots, x_d]$ and a representation of a query $q \in \mathbb{R}^d$, attention [33] computes the alignment score between q and each element x_i using a compatibility function $f(x_i, q)$. A softmax function then transforms the alignment scores $[f(x_i, q)]_{i=1}^d$ to a probability distribution $p(z \mid \mathbf{x}, q)$, where z is an indicator of which element is important to q. That is, a large $p(z = i \mid \mathbf{x}, q)$ means that x_i contributes important information to q. This attention process can be formalized as

$$\begin{aligned} \alpha &= \left[f\left(x_i, q\right)\right]_{i=1}^d, \\ p\left(z = i \mid \mathbf{x}, q\right) &= \operatorname{softmax}\left(\alpha\right). \end{aligned} \tag{1}$$

The output s_i is the weighted element according to its importance, i.e.,

$$s_i = p\left(z = i \mid \mathbf{x}, q\right) x_i. \tag{2}$$

Additive attention mechanisms [33, 34] are commonly used attention mechanisms where the compatibility function $f(\cdot)$ is parameterized by a MLP, i.e.,

$$f\left(x_i, q\right) = w^T \sigma\left(W^{(1)} x_i + W^{(2)} q\right), \tag{3}$$

where $W^{(1)} \in \mathbb{R}^{d \times d}$, $W^{(2)} \in \mathbb{R}^{d \times d}$, $w \in \mathbb{R}^d$ are learnable parameters, d is the dimension of x_i, and $\sigma(\cdot)$ is an activation function. In contrast to additive attention, multiplicative attention [35, 36] uses cosine similarity or inner product as the compatibility function for $f(x_i, q)$, i.e.,

$$f\left(x_i, q\right) = \langle W^{(1)} x_i, W^{(2)} q \rangle. \tag{4}$$

In practice, although additive attention is expensive in time cost and memory consumption, it usually achieves better empirical performance for downstream tasks.

Self-attention [37, 38] explores the importance of each feature to the entire FC given a specific task. In particular, q is removed from the common compatibility function which is formally written as the following equation:

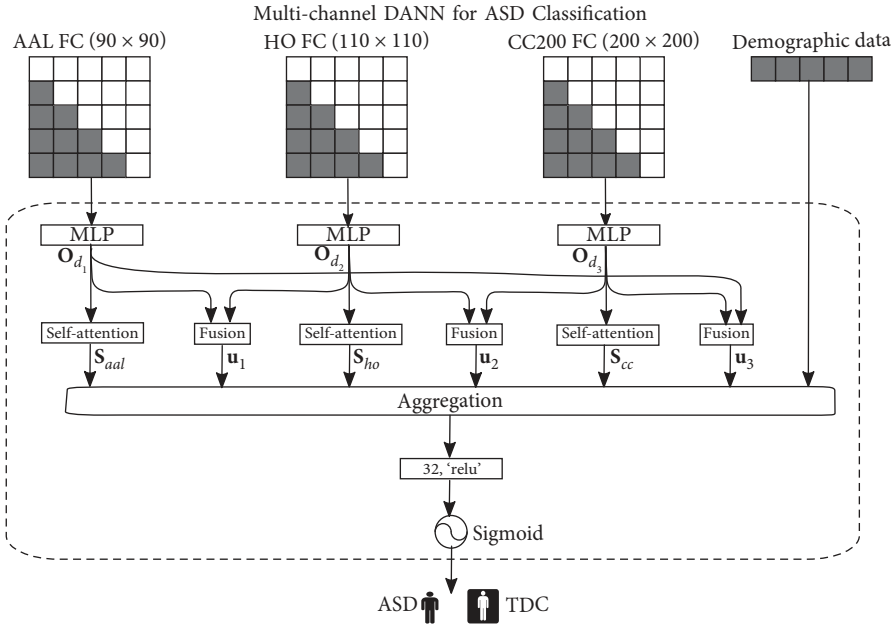

Figure 1: A DANN structure for ASD classification in this study.

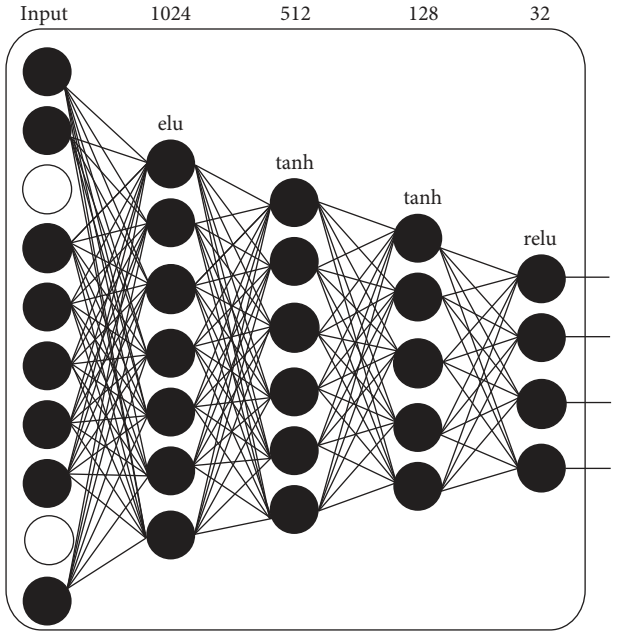

Figure 2: Detailed MLP block in DANN structure.

$$f(x_i) = w^T \sigma\left(W^{(1)} x_i\right),$$
$$\alpha = \left[f(x_i)\right]_{i=1}^{d}, \tag{5}$$
$$p(z = i \mid \mathbf{x}) = \text{softmax}(\alpha).$$

The output s_i is the weighted element according to its importance, i.e.,

$$s_i = p(z = i \mid \mathbf{x}) x_i. \tag{6}$$

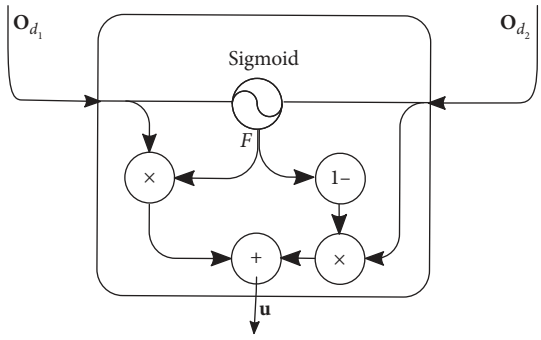

Figure 3: Detailed fusion gate in DANN structure.

2.3.4. Fusion. The fusion output u is obtained by combining the outputs of the two dense layer blocks, which can capture the correlation between the types of spaces. The combination is accomplished by a fusion gate, as shown in Figure 3, i.e.,

$$F = \text{sigmoid}\left(W^{(f_1)} o_{d_1} + W^{(f_2)} o_{d_2} + b^{(f)}\right).$$
$$u = F \odot o_{d_1} + (1 - F) \odot o_{d_2}, \tag{7}$$

where $W^{(f1)}, W^{(f2)} \in \mathbb{R}^{d_o}$, d_o is the dimension of output o_d, and $b^{(f)} \in \mathbb{R}$ are the learnable parameters of the fusion gate.

2.3.5. Aggregation. To aggregate dense layer, self-attention, and fusion into a DANN, the outputs of self-attention and fusion blocks can be concatenated, multiplied, or averaged. In our implementation, the outputs of both the self-attention blocks and the fusion blocks are concatenated, followed by a dense layer and sigmoid layer for classification:

$$l_d = \text{relu}(W_d \mathbf{v} + b_d),$$
$$\text{Comb} = \text{sigmoid}(W_c l_d + b_c), \tag{8}$$

where $\mathbf{v}$ is a vector of the combined outputs of both the self-attention blocks and the fusion blocks. $\mathbf{v} = [\mathbf{s}_{\text{aal}}, \mathbf{s}_{\text{ho}}, \mathbf{s}_{\text{cc}}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{Demo}]$ represents the concatenation of outputs $\mathbf{s}_{\text{aal}}$, $\mathbf{s}_{\text{ho}}$, $\mathbf{s}_{\text{cc}}$ from the self-attention blocks, $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$ from the fusion blocks, and **Demo** from demographic data. A sigmoid function on dense lay is then used for data classification.

3. Experiment Setup

3.1. Model Evaluation. We conducted a comprehensive evaluation in this study by employing the proposed multichannel DANN on ABIDE dataset to classify the ASD subjects from TDC subjects. Two evaluation strategies, k-fold cross validation and leave-one-site-out cross validation, were designed in our experiments. For k-fold cross validation, whole ABIDE dataset would be divided into k portions. In each repeated iteration, we randomly used one portion of the data as testing data and applied the remaining $(k-1)$ portions of the data as training data. This process would be repeated k times until all data have been tested once. For the leave-one-site-out cross validation, we separated the whole ABIDE dataset according to their data sites. We removed the SBL site from this experiment due to its small subject size ($N = 4$). This resulted in a total of 12 data sites. We randomly used data from one site as testing data and treated the remaining data from 11 data sites as training data. This is repeated 12 times until data from all sites have been evaluated as testing data. Both the k-fold cross validation and leave-one-site-out experiments were repeated 50 times to understand the variability of the results. Mean and standard deviation (SD) were calculated. Student's T-test was applied to test the difference between continuous values, and chi-square test was used for discrete values. One-way analysis of variance (ANOVA) was utilized to compare multiple conditions (i.e., multiple k-fold cross validation experiments). A p value < 0.05 was used for inferring statistical significance.

We calculated true positive (TP), false positive (FP), true negative (TN), and false negative (FN) for the classification by comparing the classified labels and gold-standard labels. Then, we calculated accuracy, sensitivity, precision, and F-score by

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$
$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{9}$$
$$F - \text{score} = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}},$$
$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

3.2. Peer Machine Learning Models. To compare our multichannel DANN with existing machine learning models, we also implemented random forest (RF), support vector machine (SVM) models, and multichannel DNN. Each model was designed to take multimodality data as inputs.

3.2.1. Random Forest (RF). RF is one of the classic ensemble learning methods by learning multiple decision trees to improve classification performance and control overfitting. The number of trees in the forest was optimized from empirical values [20, 40, 60, 80, 100]. We set the maximal depth of the tree as 10.

3.2.2. Support Vector Machine (SVM). A SVM model was developed to perform ASD classification by using vectorized FC features. We applied a linear kernel and searched the margin penalty with empirical values [0.2, 0.4, 0.6, 0.8, 1.0].

3.2.3. Deep Neural Networks (DNNs). In terms of existing deep learning model, we compared our model with a DNN model developed previously for ASD classification [26]. In brief, the compared existing DNN model is a 5-layer DNN, with input number of nodes in input layer, followed by 1024, 512, 128, and 32 nodes in hidden layers, and the output layer contains two output units. A cross entropy loss function was adopted. Learning rate was set as 0.0001. 10 epochs were applied to ensure the convergence of the model.

3.3. Developmental Environment. The proposed DANN and peer machine learning models were implemented in the Python 3.7 environment. To build the deep learning related models, we applied Keras (2.2.4) package with TensorFlow (1.13.1) backend. For the traditional models, we adopted the models from Sklearn 0.20 [39]. Statistical analyses were performed using Matlab 2019b.

All the experiments were conducted on a workstation with 10 cores of Intel Core i9 CPU and 64 GB RAM. Due to the high computation cost of deep learning algorithm, we configured one GPU (Nvidia TITAN Xp, 12 GB RAM) to accelerate the training speed of the models.

4. Results and Discussion

4.1. Performance Comparison on the Whole ABIDE Dataset. We first compared the ASD classification performance of the proposed multichannel DANN model and multiple peer machine learning models, including RF, SVM, and multichannel DNN. The results were calculated based on 50 repeats of 10-fold cross validation experiments by using the entire ABIDE dataset. The mean and SD of the performance metrics are listed in Table 2. The proposed multichannel DANN exhibited a significantly higher accuracy than multichannel DNN ($p = 0.01$), SVM ($p = 0.014$), and RF ($p = 0.008$) models. Similarly, the multichannel DANN also had better F-score than multichannel DNN ($p = 0.004$), SVM ($p < 0.001$), and RF ($p < 0.001$) models. The sensitivity of the multichannel

Table 2: Comparison of random forest (RF), support vector machine (SVM), multichannel deep neural network (DNN), and multichannel deep attention neural network (DANN) classifiers trained using 10-fold cross validation on the entire dataset.

Method	Accuracy	Sensitivity	Precision	F-Score	Specificity
RF	0.659 ± 0.018	0.689 ± 0.106	0.656 ± 0.012	0.671 ± 0.023	0.628 ± 0.081
SVM	0.693 ± 0.059	0.713 ± 0.059	0.696 ± 0.072	0.702 ± 0.048	0.673 ± 0.113
Multichannel DNN	0.707 ± 0.027	0.673 ± 0.088	$\mathbf{0.740 \pm 0.106}$	0.718 ± 0.060	0.700 ± 0.067
Multichannel DANN	$\mathbf{0.732 \pm 0.024}$	$\mathbf{0.745 \pm 0.115}$	0.730 ± 0.053	$\mathbf{0.736 \pm 0.042}$	$\mathbf{0.717 \pm 0.101}$

All data are mean and standard deviation. The highest metrics were marked as bold.

DANN was significantly higher than that of multichannel DNN ($p = 0.009$), SVM ($p = 0.015$), and RF ($p = 0.005$) models. The specificity of the multichannel DANN was significantly higher than that of SVM ($p = 0.004$) and RF ($p < 0.001$) models but was not significantly better than multichannel DNN ($p = 0.082$). Since the multichannel DNN had a relatively lower sensitivity (0.673), it achieved the best mean precision in our experiments. No significant difference ($p = 0.219$) was found between multichannel DNN and DANN on precision. The multichannel DANN model still exhibited higher precision than SVM ($p = 0.003$) and RF ($p < 0.001$). Overall, the proposed multichannel DANN achieved improved ASD classification accuracy, sensitivity, *F*-score, and specificity among compared machine learning models, while the multichannel DNN had the highest precision.

Inspiringly, the proposed multichannel DANN significantly outperformed multichannel DNN on four of five performance metrics, increasing mean accuracy by 0.025, sensitivity by 0.072, *F*-score by 0.018, and specificity by 0.017. Although no significance was found, the precision of the proposed approach is slightly lower than multichannel DNN by 0.01. The attention mechanism in our model, as the name implies, aids the deep learning model to make choices about which features it should pay attention. Our model can allocate attention by adjusting the weights they assign to individual FC features. This process can decide which FC features are more important than others in terms of the ASD classification task. In another word, it optimizes the feature selection during the learning of a deep learning model. The improved performance of DANN over DNN demonstrated the validity of the attention mechanism. The results in Table 2 also showed that multichannel DANN achieved significantly improved performance, compared to traditional models SVM and RF. This is consistent with multiple previous ASD classification studies [26, 27]. The improvement was likely due to a combination of attention mechanism and the superior capability of deep learning model on complex data patterns, such as FC features.

4.2. Leave-One-Site-Out Cross Validation of Multichannel DANN. To test the generalizability of the proposed model on unseen data from different data sites, we performed a leave-one-site-out cross validation. Similar to *k*-fold cross validation, we reserved data from one data site as testing data and trained our model by using all data from the rest of the 11 data sites. But, since the training data were the same across all repeats, the performances have much smaller variations than *k*-fold cross validation. Table 3 shows the classification performance of our model and the size of subjects for each data site.

In the NYU data site that contains the largest sample size, our model achieved an accuracy of 0.709 ± 0.019, sensitivity of 0.720 ± 0.086, the precision of 0.758 ± 0.127, *F*-score of 0.738 ± 0.069, and specificity of 0.689 ± 0.072. When examining data sites with more than 40 subjects, we found that our model achieved the highest accuracy (0.803 ± 0.045) on the USM site and the best *F*-score (0.745 ± 0.052) on the UCLA site. These two sites contain nearly 100 subjects, so the results are very informative. We also noted that the lowest accuracy our model returned was 0.684 ± 0.026 from UM site, suggesting that the data here may have variability that is different from other sites. Overall, our model reached a mean accuracy of 0.713 ± 0.022 and mean *F*-score 0.707 ± 0.043. This was significantly lower than accuracy ($p = 0.002$) and *F*-score ($p < 0.001$) from the cross validation results in Table 2, indicating a large data variability among different data sites.

4.3. Robustness of Multichannel DANN on Varying Data Split Schemes. Next, the robustness of our DANN was further tested using varying *k*-fold cross validation. A classification model that is not robust may appear to perform very differently with different *k*. Figure 4 shows plots of the accuracy, sensitivity, precision, *F*-score, and specificity of the proposed DANN over *k*-fold cross validation strategies ($k = [6, 7, 8, 9, 10]$). Using one-way ANOVA, the proposed DANN exhibited no significantly different performance across varying *k*-fold experiments ($p = 0.082$), indicating the robustness of the proposed multichannel DANN model.

4.4. Impact of Data Modality on the Classification Performance. At the end, we set to test the performance of the multichannel DANN when different data modalities are used for ASD classification. All results were based on 50 repeats of 10-fold cross validation experiment. Table 4 lists the performance of multichannel DANN on varying combinations of FC data (marked as AAL, HO, and CC200) and PC data (marked as Demo). The upper part of Table 4 contains results based on both FC and PC data, while the lower part of the table focuses on FC data only. The combined FC and PC data (AAL + HO + CC + Demo) had a better accuracy ($p = 0.011$), sensitivity ($p = 0.039$), and specificity ($p = 0.025$) than FC data alone (AAL + HO + CC), while no significant differences were

Table 3: Leave-one-site-out cross validation results using multichannel DANN.

Site-out	Size	Accuracy	Sensitivity	Precision	*F*-score	Specificity
TRINITY	46	0.696 ± 0.012	0.640 ± 0.012	0.762 ± 0.036	0.696 ± 0.004	0.679 ± 0.070
YALE	56	0.696 ± 0.025	0.679 ± 0.029	0.714 ± 0.032	0.691 ± 0.034	0.682 ± 0.065
STANFORD	39	0.615 ± 0.018	0.350 ± 0.025	0.778 ± 0.039	0.483 ± 0.015	0.685 ± 0.032
SDSU	36	0.694 ± 0.024	0.727 ± 0.095	0.762 ± 0.072	0.744 ± 0.059	0.705 ± 0.067
CALTECH	36	0.667 ± 0.029	0.556 ± 0.016	0.714 ± 0.029	0.625 ± 0.015	0.693 ± 0.038
UCLA	98	0.755 ± 0.015	0.795 ± 0.017	0.700 ± 0.009	0.745 ± 0.012	0.701 ± 0.019
CMU	27	0.630 ± 0.019	0.692 ± 0.044	0.600 ± 0.037	0.643 ± 0.044	0.684 ± 0.035
USM	71	0.803 ± 0.015	0.560 ± 0.028	0.824 ± 0.034	0.667 ± 0.029	0.685 ± 0.038
NYU	175	0.709 ± 0.019	0.720 ± 0.026	0.758 ± 0.027	0.738 ± 0.039	0.689 ± 0.022
PITT	56	0.696 ± 0.022	0.778 ± 0.023	0.656 ± 0.002	0.712 ± 0.027	0.717 ± 0.013
LEUVEN	29	0.621 ± 0.017	1.000 ± 0.017	0.577 ± 0.027	0.732 ± 0.028	0.674 ± 0.022
UM	126	0.684 ± 0.026	0.761 ± 0.008	0.675 ± 0.009	0.715 ± 0.008	0.671 ± 0.012
Mean	62	0.713 ± 0.022	0.712 ± 0.081	0.731 ± 0.087	0.707 ± 0.043	0.713 ± 0.057

All data are mean and standard deviation.

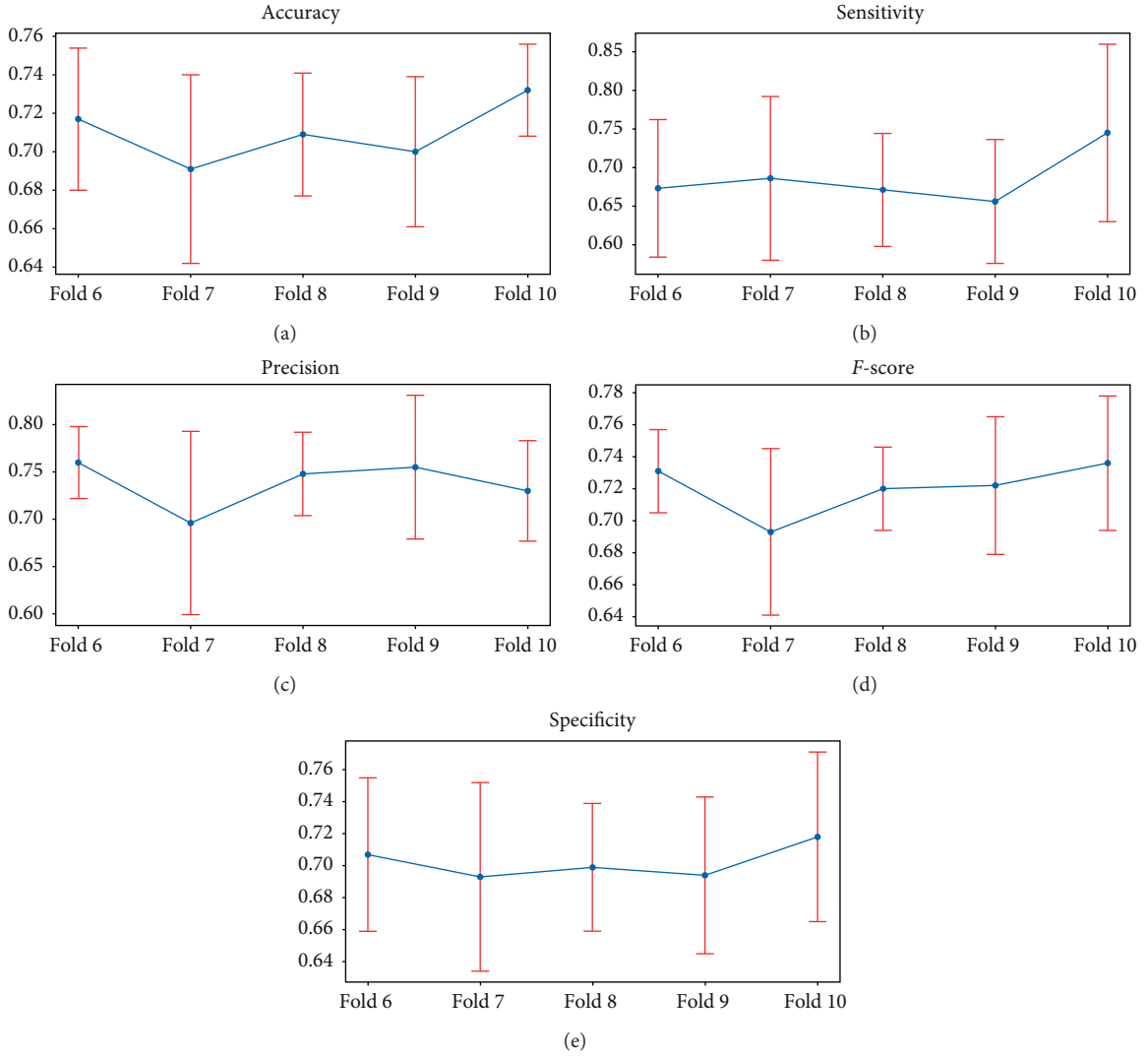

Figure 4: Performance of multichannel DANN over varying data split schemes with k-fold cross validation strategies ($k = [6, 7, 8, 9, 10]$). Mean and standard deviation are displayed.

Table 4: Comparison of multichannel DANN on different data combinations using 10-fold cross validation on the entire dataset.

Data	Accuracy	Sensitivity	Precision	*F*-score	Specificity
AAL + HO + CC + Demo	0.732 ± 0.024	0.745 ± 0.115	0.730 ± 0.053	0.736 ± 0.042	0.717 ± 0.101
AAL + HO + Demo	0.700 ± 0.035	0.698 ± 0.068	0.701 ± 0.035	0.702 ± 0.004	0.673 ± 0.401
AAL + CC + Demo	0.703 ± 0.009	0.721 ± 0.084	0.686 ± 0.071	0.699 ± 0.067	0.697 ± 0.060
HO + CC + Demo	0.691 ± 0.018	0.696 ± 0.098	0.686 ± 0.106	0.690 ± 0.054	0.687 ± 0.065
AAL + Demo	0.666 ± 0.002	0.683 ± 0.080	0.650 ± 0.071	0.659 ± 0.006	0.679 ± 0.031
HO + Demo	0.689 ± 0.027	0.681 ± 0.078	0.696 ± 0.106	0.691 ± 0.011	0.685 ± 0.057
CC + Demo	0.692 ± 0.053	0.703 ± 0.092	0.681 ± 0.141	0.689 ± 0.043	0.704 ± 0.055
AAL + HO + CC	0.720 ± 0.062	0.696 ± 0.097	0.738 ± 0.212	0.724 ± 0.078	0.695 ± 0.056
AAL + HO	0.684 ± 0.027	0.636 ± 0.084	0.730 ± 0.018	0.699 ± 0.035	0.673 ± 0.050
AAL + CC	0.695 ± 0.018	0.683 ± 0.083	0.706 ± 0.124	0.700 ± 0.030	0.683 ± 0.052
HO + CC	0.688 ± 0.027	0.666 ± 0.086	0.711 ± 0.053	0.697 ± 0.020	0.683 ± 0.055
AAL	0.658 ± 0.009	0.641 ± 0.078	0.674 ± 0.141	0.663 ± 0.035	0.658 ± 0.032
HO	0.679 ± 0.009	0.683 ± 0.065	0.674 ± 0.071	0.677 ± 0.007	0.691 ± 0.029
CC	0.682 ± 0.005	0.651 ± 0.071	0.713 ± 0.106	0.693 ± 0.034	0.677 ± 0.046

AAL: AAL atlas-based FC; HO: HO atlas-based FC; CC: CC200 atlas-based FC; Demo: PC data. All data are mean and standard deviation.

observed on precision ($p = 0.231$) and *F*-score ($p = 0.347$). This demonstrated the predictive power of PC data.

Without PC data, our model achieved the highest performance by combining FC from all three brain atlases. This suggests that brain connected data from different atlases may have complementary information so as to assist the ASD classification. Interestingly, the model using CC200 FC data (marked as CC in the table) performed better than FC data derived from AAL ($p = 0.012$) and HO ($p = 0.023$). It is likely because that CC200 atlas is constructed from rs-fMRI data, representing a brain functional parcellation.

5. Conclusion

In summary, we developed a multichannel DANN model by applying the state-of-the-art attention mechanism-based deep learning techniques for automated diagnosis of ASD. The *k*-fold cross validation experiments have shown that our multichannel DANN achieved an accuracy of 0.732, outperforming multiple peer machine learning models. The results of the leave-one-site-out cross validation experiments showed promise for our model to be applied to clinical data with unseen variations. The experiments using varying combinations of data modalities demonstrated discriminative power of individual data modalities such as brain functional connectome and PC data. This suggests a future direction of combining additional data modalities to move the machine learning applications towards clinical usage of ASD computer-aided diagnosis tools. One limitation of the current work is that the selected cohort is in the adolescent and young adult population, which limits the generalizability of the model, since the ASD diagnosis was performed much earlier. In the future study, we would retrain the model with additional data from a wider age range of population.

Data Availability

The dataset used to support the findings of this study is available in http://fcon_1000.projects.nitrc.org/indi/abide/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the Beijing Education Commission Research Project of China under grant no. KM201911232004, National Natural Science Foundation of China under grant no. 61672105, and National Key Research and Development Program of China under grant no. 2018YFB1004100.

References

[1] J. Baio, *Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years-Autism and Developmental Disabilities Monitoring Network, 11 Sites*, Centers for Disease Control and Prevention, Atlanta, GA, USA, 2010.

[2] L. Tonello, L. Giacobbi, A. Pettenon et al., "Crisis behavior in autism spectrum disorders: a self-organized criticality approach," *Complexity*, vol. 2018, pp. 1–7, 2018.

[3] E. Simonoff, A. Pickles, T. Charman, S. Chandler, T. Loucas, and G. Baird, "Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 47, no. 8, pp. 921–929, 2008.

[4] S. Goldstein and S. Ozonoff, *Assessment of Autism Spectrum Disorder*, Guilford Publications, New York, NY, USA, 2018.

[5] I. Riquelme, S. M. Hatem, and P. Montoya, "Abnormal pressure pain, touch sensitivity, proprioception, and manual dexterity in children with autism spectrum disorders," *Neural Plasticity*, vol. 2016, Article ID 1723401, 9 pages, 2016.

[6] R. Djemal, K. AlSharabi, S. Ibrahim, and A. Alsuwailem, "EEG-based computer aided diagnosis of autism spectrum disorder using wavelet, entropy, and ann," *BioMed Research International*, vol. 2017, Article ID 9816591, 9 pages, 2017.

[7] K. B. Schauder and L. Bennetto, "Toward an interdisciplinary understanding of sensory dysfunction in autism spectrum disorder: an integration of the neural and symptom literatures," *Frontiers in Neuroscience*, vol. 10, p. 268, 2016.

[8] N. Newbutt, C. Sung, H. J. Kuo, and M. J. Leahy, "The acceptance, challenges, and future applications of wearable technology and virtual reality to support people with autism spectrum disorders," in *Recent Advances in Technologies for Inclusive Well-Being*, pp. 221–241, Springer, Berlin, Germany, 2017.

[9] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*, American Psychiatric Association Publishing, GA, USA, 2013.

[10] M. Galliver, E. Gowling, W. Farr, A. Gain, and I. Male, "Cost of assessing a child for possible autism spectrum disorder? an observational study of current practice in child development centres in the UK," *BMJ Paediatrics Open*, vol. 1, no. 1, Article ID e000052, 2017.

[11] K. K. Hyde, M. N. Novack, N. LaHaye et al., "Applications of supervised machine learning in autism spectrum disorder research: a review," *Review Journal of Autism and Developmental Disorders*, vol. 6, no. 2, pp. 128–146, 2019.

[12] D. Gil, M. Johnsson, H. Mora, and J. Szymanski, "Advances in architectures, big data, and machine learning techniques for complex Internet of things systems," *Complexity*, vol. 2019, Article ID 4184708, 3 pages, 2019.

[13] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[14] Ç. Uğur, H. Tunca, E. Sekmen et al., "A comparative study of the oxidative stress indices of children with autism and healthy children," *Anatolian Journal of Psychiatry*, vol. 19, no. 3, 2018.

[15] C. Li, H. Zhou, T. Wang et al., "Performance of the autism spectrum rating scale and social responsiveness scale in identifying autism spectrum disorder among cases of intellectual disability," *Neuroscience Bulletin*, vol. 34, no. 6, pp. 972–980, 2018.

[16] R. L. Hansen, N. J. Blum, A. Gaham et al., "Diagnosis of autism spectrum disorder by developmental-behavioral pediatricians in academic centers: a DBPNet study," *Pediatrics*, vol. 137, no. 2, pp. S79–S89, 2016.

[17] J. N. Constantino and C. P. Gruber, *Social Responsiveness Scale (SRS), Western Psychological Services*, Springer, New York, NY, USA, 2007.

[18] T. Charman, A. Pickles, E. Simonoff, S. Chandler, T. Loucas, and G. Baird, "IQ in children with autism spectrum disorders: data from the special needs and autism project (SNAP)," *Psychological Medicine*, vol. 41, no. 3, pp. 619–627, 2011.

[19] S. L. Bishop, J. Richler, and C. Lord, "Association between restricted and repetitive behaviors and nonverbal IQ in children with autism spectrum disorders," *Child Neuropsychology*, vol. 12, no. 4-5, pp. 247–267, 2006.

[20] 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, Article ID e0166934, 2016.

[21] 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, Article ID e0194856, 2018.

[22] G. J. Katuwal, S. A. Baum, N. D. Cahill, and M. M. Andrew, "Divide and conquer: sub-grouping of asd improves ASD detection based on brain morphometry," *PLoS One*, vol. 11, no. 4, Article ID e0153331, 2016.

[23] C. Zu, Y. Gao, B. Munsell et al., "Identifying disease-related subnetwork connectome biomarkers by sparse hypergraph learning," *Brain Imaging and Behavior*, vol. 13, no. 4, pp. 879–892, 2018.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Lake Tahoe, NV, USA, 2012.

[26] 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*, vol. 17, pp. 16–23, 2018.

[27] Y. Kong, J. Gao, Y. Xu, Y. Pan, J. Wang, and J. Liu, "Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier," *Neurocomputing*, vol. 324, pp. 63–68, 2019.

[28] T. Eslami, V. Mirjalili, A. Fong, A. Laird, and F. Saeed, "ASD-diagnet: a hybrid learning approach for detection of autism spectrum disorder using FMRI data," 2019, http://arxiv.org/abs/1904.07577.

[29] J. Guo, K. Yang, H. Liu et al., "A stacked sparse autoencoder-based detector for automatic identification of neuromagnetic high frequency oscillations in epilepsy," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2474–2482, 2018.

[30] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. G. Kame, "Knowledge-based attention model for diagnosis prediction in healthcare," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 743–752, ACM, Torino, Italy, October 2018.

[31] X. Peng, G. Long, T. Shen, S. Wang, J. Jiang, and M. Blumenstein, "Temporal self-attention network for medical concept embedding," 2019, http://arxiv.org/abs/1909.06886.

[32] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: directional self-attention network for RNN/CNN-free language understanding," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, February 2018.

[33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, http://arxiv.org/abs/1409.0473.

[34] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," 2015, http://arxiv.org/abs/1503.02364.

[35] S. Sukhbaatar, J. Weston, R. Fergus et al., "End-to-end memory networks," in *Proceedings of the Conference on Neural Information Processing Systems*, Montreal, Canada, December 2015.

[36] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," 2015, http://arxiv.org/abs/1509.00685.

[37] Z. Lin, M. Feng, C. Nogueira dos Santos et al., "A structured self-attentive sentence embedding," 2017, http://arxiv.org/abs/1703.03130.

[38] Y. Liu, C. Sun, L. Lin, and X. Wang, "Learning natural language inference using bidirectional LSTM model and inner-attention," 2016, http://arxiv.org/abs/1605.09090.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.