Article:
Kunda, M., Zhou, S., Gong, G. et al. (2022) Improving multi-site autism classification via site-dependence minimization and second-order functional connectivity. IEEE Transactions on Medical Imaging, 42 (1). pp. 55-65. ISSN: 0278-0062

https://doi.org/10.1109/tmi.2022.3203899

Improving Multi-Site Autism Classification via Site-Dependence Minimization and Second-Order Functional Connectivity

Mwiza Kunda, Shuo Zhou, Gaolang Gong, and Haiping Lu, *Senior Member, IEEE*

***Abstract*— Machine learning has been widely used to develop classification models for autism spectrum disorder (ASD) using neuroimaging data. Recently, studies have shifted towards using large multi-site neuroimaging datasets to boost the clinical applicability and statistical power of results. However, the classification performance is hindered by the heterogeneous nature of agglomerative datasets. In this paper, we propose new methods for multi-site autism classification using the Autism Brain Imaging Data Exchange (ABIDE) dataset. We firstly propose a new second-order measure of functional connectivity (FC) named as Tangent Pearson embedding to extract better features for classification. Then we assess the statistical dependence between acquisition sites and FC features, and take a domain adaptation approach to minimize the site dependence of FC features to improve classification. Our analysis shows that 1) statistical dependence between site and FC features is statistically significant at the 5% level, and 2) extracting second-order features from neuroimaging data and minimizing their site dependence can improve over state-of-the-art classification results, achieving a classification accuracy of 73%. The code is available at https://github.com/kundaMwiza/fMRI-site-adaptation.**

***Index Terms*— Data heterogeneity, domain adaptation, fMRI, autism spectrum disorders, functional connectivity.**

I. Introduction

Autism spectrum disorder (ASD) refers to a lifelong neurodevelopmental disorder characterised by a wide range of symptoms, skills and levels of disability, such as deficits in social communication, interaction and the presentation of repetitive patterns of behaviour or restricted interests [1]. Autism diagnosis is challenged by the significant behavioural heterogeneity and wide array of neuroanatomical abnormalities exhibited between patients with autism [2], [3].

Non-invasive brain imaging techniques such as magnetic resonance imaging (MRI) have been used to discover structural or functional differences between ASD and typical control (TC) subjects. In particular, resting-state functional MRI (rs-fMRI) has achieved promising results when utilized with machine learning (ML) models for classifying ASD and TC subjects [4]. However, the clinical generalizability of most studies using rs-fMRI data for autism classification is debatable since the sample sizes used are small, unlikely to cover a wide spectrum of autism and its heterogeneity [5]. These small sample sizes are due to the time and cost constraint imposed upon *single-site* studies acquiring rs-fMRI using a single fMRI scanner and subject acquisition protocol.

To improve the statistical power and generalizability of neuroimaging studies, the Autism Brain Imaging Data Exchange (ABIDE) initiative has aggregated data from multiple sites across the world, creating datasets much larger than those used in single-site studies [6]. The ABIDE dataset is composed of rs-fMRI and phenotypic data from 20 different international sites, leading to a heterogeneous sample of over 1000 ASD and TC subjects. While it presents a great potential for the extraction of functional biomarkers for autism classification, its multi-site and multi-protocol aspects bring along significant patient heterogeneity, statistical noise and experimental differences in the rs-fMRI data, making the classification task much more challenging [7]. Recent works have employed different ML methods, such as recurrent neural networks (RNN), graph convolutional neural networks (GCN) and autoencoders [8]–[12]. However, despite the complexity in patterns that these methods can generally capture, the difference in their top classification results on ABIDE fall less than 1%, with the highest achieved accuracy being 70.4% [11].

This paper investigates two research questions that can potentially improve multi-site autism classification.

- **Between-site heterogeneity:** *how can we effectively account for the experimental differences in the ABIDE rs-fMRI data?* Previous studies have reported that between-site heterogeneity arising from the use of different fMRI scanner types and experimental settings has an impact on the image properties of rs-fMRI data, and that this consequently impacts any rs-fMRI analysis [14], [15].
- **Discriminative features:** *can we design new rs-fMRI features for better autism classification?* As pointed out above, powerful and complex ML methods such as RNN, GCN, and denoising autoencoders give similar top classification performance of less than 1% difference, whether directly using the time series or employing *functional connectivity* (FC) features.

This work was supported in part by the UK Engineering and Physical Sciences Research Council under Grant EP/R014507/1, and the National Natural Science Foundation of China under Grant 81671772.

Mwiza Kunda was with the Department of Computer Science, University of Sheffield, S1 4DP, U.K. (e-mail: mwizakunda@gmail.com).

Shuo Zhou and Haiping Lu are with the Department of Computer Science, University of Sheffield, S1 4DP, U.K. (e-mail: shuo.zhou@sheffield.ac.uk; h.lu@sheffield.ac.uk).

Gaolang Gong is with State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, 100875, China (e-mail: gaolang.gong@bnu.edu.cn).

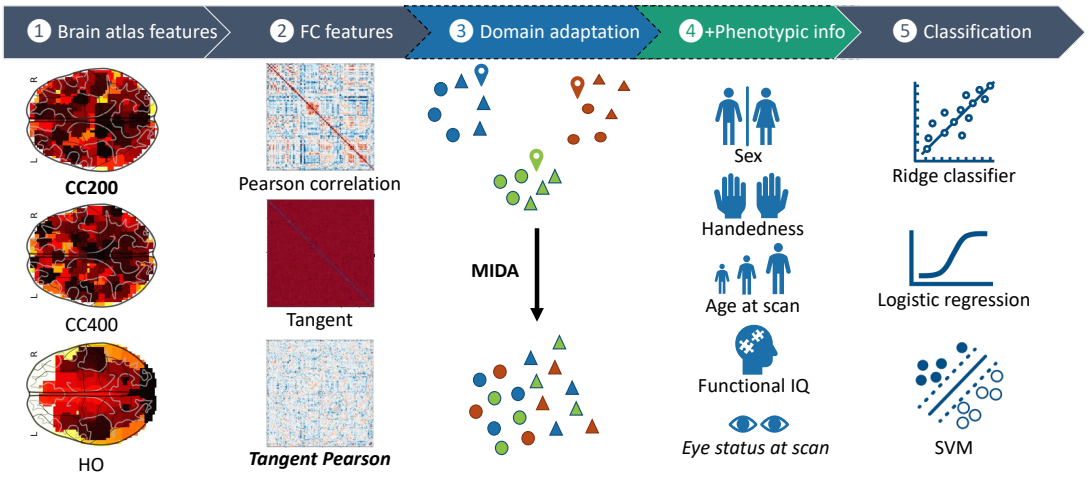

Fig. 1. The pipeline for domain adaptation on FC features, covering all models studied in this paper. Steps 1, 2 and 5 are compulsory, whilst 3 and 4 are optional. Step 3 performs maximum independence domain adaptation (MIDA) [13] on FC features and step 4 supplements subject feature vectors with phenotypic information. The phenotype 'eye status at scan' is site-specific so it is excluded when MIDA is used.

Domain adaptation methods operate on datasets from different sources with mismatched distributions to find a new latent space where the data is homogeneous, or source invariant [16], [17]. In the context of this study, this corresponds to aligning the rs-fMRI data so that there is *independence* between the data and acquisition sites. Recently, Moradi et al. [18] proposed a domain adaptation approach to correct site heterogeneity for the estimation of symptom severity in autism using ABIDE. Their severity score predictions were markedly better than those from models without domain adaptation. However, their study was limited to only 156 subjects from 4 of the 20 sites and they did not tackle the classification problem. Another recent work [19] takes a low-rank representation approach for multi-site domain adaption but their study was limited to 468 subjects from 5 of the 20 sites. In contrast, our study focuses on the technical challenge of assessing and targeting the site heterogeneity in all 20 sites to improve autism classification.

Functional connectivity (FC) measures are important features in ASD classification [20]. Two popular FC measures are: 1) the Pearson correlation measures the coupling between pairs of regions of interest (ROIs), and 2) the more recent tangent embedding parameterization of the covariance matrix captures the FC differences between a single subject and a group [21]. In this paper, we explore a new perspective: *for any two ROIs, are they functionally connected to other brain regions in the same way?* This inspires us to propose a new second-order FC measure that *jointly* considers the FC of individual ROIs.

In this study, we analyzed the rs-fMRI data of 1035 subjects from all 20 ABIDE sites to improve multi-site autism classification. We focused on constructing a new second-order FC measure and evaluating the impact of minimizing their dependence on the acquisition sites for autism classification. The main contributions are threefold:

- We proposed a new second-order FC measure, *Tangent Pearson (TP) embedding* to extract more discriminative features for multi-site autism classification, outperforming two popular FC measures on the whole.
- We assessed the statistical significance of the dependence between FC features and acquisition sites, showing significance at the 5% level and motivating us to design models that correct for between-site heterogeneity.
- We took a domain adaptation approach to minimize the dependence between acquisition sites and FC features. Combining with the TP measure and phenotypic information, this approach improved autism classification on ABIDE, yielding state of the art results.

II. Material and Methods

Figure 1 gives an overview of the pipeline for studying domain adaptation on FC features. It shows the steps involved in specifying various models. Step 3 is the proposed domain adaptation step optional in the pipeline and it is used to extract site-independent features from the FC data of step 2 in an *unsupervised* way. The impact of using such features can then be compared with models that do not use step 3. Likewise, the including of phenotypic information is optional.

A. ABIDE database: rs-fMRI and phenotypic data

This study focuses on the ABIDE database, which is composed of MRI and phenotypic data collected from 20 sites around the world. We included rs-fMRI and phenotypic data from 505 ASD and 530 TC individuals, yielding a sample of 1035 subjects. This sample of subjects is the same as that used in [9], which differs from the 871 subjects used in [7], [11] due to their use of image quality control measures upon the full database. We opted for such a large sample to increase the likelihood of detecting site effects from individual sites, even though a larger sample presents the challenge of a greater level of heterogeneity expressed in subjects.

ABIDE provides a range of phenotypic information, including factors such as sex, age, full IQ (FIQ) test scores and handedness (left, right or ambidextrous). In particular, the type of fMRI scanner and length of individual fMRI scans varied across sites, giving rise to the apparent heterogeneity. Table I

TABLE I

Phenotypic and experimental variation across ABIDE sites. FOR QUANTITATIVE VARIABLES, THE STANDALONE VALUES REPRESENT THE OBSERVED MEANS, SD REPRESENTS THE STANDARD DEVIATION. M: MALE, F: FEMALE, L/R: LEFT/RIGHT HANDINESS.

Site ID	Scanner	Handedness (L/R)	Eye Status	Sex (M/F)	Age (SD)	FIQ (SD)	Scan Time (SD)
CALTECH	SIEMENS Trio	9/28	Closed	29/8	27.72 (10.45)	111.16 (11.53)	146.0 (0.0)
CMU	SIEMENS Verio	3/24	Closed	21/6	26.59 (5.69)	114.56 (10.54)	273.26 (42.46)
KKI	Philips Achieva	8/40	Open	36/12	10.01 (1.27)	106.17 (15.0)	148.5 (9.36)
LEUVEN_1	Philips INTERA	2/27	Open	29/0	22.59 (3.55)	112.21 (13.03)	246.0 (0.0)
LEUVEN_2	Philips INTERA	4/28	Closed	26/8	14.09 (1.38)	100.0 (0.0)	246.0 (0.0)
MAX_MUN	SIEMENS Verio	2/50	Closed	48/4	25.31 (11.88)	110.52 (11.73)	140.62 (37.28)
NYU	SIEMENS Allegra	0/175	Open	139/36	15.26 (6.57)	110.51 (14.99)	176.0 (0.0)
OHSU	SIEMENS Trio	1/25	Open	26/0	10.71 (1.79)	110.6 (16.81)	78.0 (0.0)
OLIN	SIEMENS Allegra	6/28	Open	29/5	16.59 (3.47)	112.41 (17.01)	206.0 (0.0)
PITT	SIEMENS Allegra	4/52	Closed	48/8	18.94 (6.93)	110.18 (12.24)	196.0 (0.0)
SBL	Philips Intera	1/29	Closed	30/0	34.37 (8.6)	101.53 (6.15)	196.0 (0.0)
SDSU	GE MR750	4/32	Open	29/7	14.41 (1.84)	109.36 (13.77)	176.0 (0.0)
STANFORD	GE Signa	6/33	Closed	31/8	9.98 (1.59)	111.41 (15.56)	209.77 (30.02)
TRINITY	Philips Achieva	0/47	Closed	47/0	16.96 (3.47)	109.96 (13.75)	146.0 (0.0)
UCLA_1	SIEMENS Trio	6/66	Open	62/10	13.19 (2.4)	103.46 (12.02)	116.0 (0.0)
UCLA_2	SIEMENS Trio	4/22	Open	24/2	12.49 (1.53)	102.04 (14.92)	116.0 (0.0)
UM_1	GE Signa	14/92	Open	81/25	13.4 (2.88)	104.96 (14.11)	296.0 (0.0)
UM_2	GE Signa	3/31	Open	32/2	16.01 (3.36)	112.26 (10.85)	296.0 (0.0)
USM	SIEMENS Trio	0/71	Open	71/0	22.69 (8.34)	105.23 (17.65)	235.94 (0.47)
YALE	SIEMENS Trio	10/46	Open	40/16	12.71 (2.88)	99.77 (20.12)	196.0 (0.0)

gives a summary of each site with respect to key experimental protocols and phenotypic information. It shows that the type of fMRI scanners and length of individual fMRI scans varied across sites, giving rise to the apparent heterogeneity.

To compare against the state-of-the-art (SOTA) methods [7], [9], [11], we used the same pre-processed fMRI data from ABIDE (http://preprocessed-connectomes-project.org/abide/).

B. Step 1: Brain atlas features

Studies on rs-fMRI typically define brain regions of interest (ROIs) rather than operating on individual voxels. These ROIs represent the aggregation (e.g., averaging) of the rs-fMRI time series data of individual voxels so that the number of ROIs is significantly less than the number of voxels.

We chose the Craddock 200 (CC200) brain atlas [22] due to its robust performance in previous studies on ABIDE [8], [9], [11]. CC200 has 200 ROIs derived from the clustering of spatially close voxels. We also considered two additional atlases to assess the impact of using a different brain parcellation: 1) Harvard Oxford (HO), a structural atlas with 110 ROIs based on anatomical landmarks from 40 sMRI scans [23], and 2) Craddock 400 (CC400), an atlas with 392 ROIs computed in a similar way to the CC200 atlas. For these three atlases, the representative time series of an ROI was derived by averaging the rs-fMRI time series of voxels associated with the ROI.

C. Step 2: Functional connectivity features

FC features are usually extracted between pairs of ROIs based on the raw time series data. They estimate the fluctuating coupling of brain regions with respect to time, so that we can train predictive models for classification based on differences in brain region coupling. We first consider two SOTA FC measures as baselines: 1) the Fisher transformed *Pearson*'s correlation coefficient used in [11], which gives a measure of coupling between pairs of ROIs by computing the correlation between their time series, and 2) the *tangent embedding* parameterization of the covariance matrix proposed in [21], which captures the deviation of each subject covariance matrix from the group mean covariance matrix and outperforms many other FC measures in [7]. These two FC measures have achieved SOTA autism classification performance on ABIDE.

Proposed second-order FC measure. As reviewed in Sec. I, various ML methods including RNN and GCN have been applied on the above FC features for multi-site autism classification. However, the resulting top classification accuracies differ by less than 1%. This makes us question whether such simple FC measures can capture well the complexity of brain networks. Thus, it motivates us to go a step further than the two standard FC measures and construct a new *second-order* FC measure in the following. The Pearson correlation coefficient gives a measure of the coupling between ROIs pairs irrespective of any other ROIs. We propose to also quantify the relationship that two ROIs have with respect to all other ROIs. That is, we want to examine the following question: *for any two ROIs, are they functionally connected to other brain regions in the same way?*

Given a set of R ROIs and a corresponding FC matrix, $\mathbf{M}$ ($R \times R$), each row of $\mathbf{M}$, $\mathbf{M}_i, i \in \{1, \ldots, R\}$, gives the measure of FC between region i and all other regions. So given any two regions i and j, the *second-order* measure of interest can be computed by measuring the similarity between $\mathbf{M}_i$ and $\mathbf{M}_j$. We propose to capture this second-order measure by firstly computing the Pearson correlation coefficient between pairs of ROI time series, and then computing the covariance of the resulting connectivity profiles from all regions (e.g. $\mathbf{M}_i$ and $\mathbf{M}_j$). The Pearson's correlation coefficient at the first step is powerful in capturing first-order FC measures. Denoting a correlation matrix as $\mathbf{R}$, it is given by

$$\mathbf{R} = \mathbf{\Sigma}(1/\mathrm{diag}(\mathbf{\Sigma})^{\frac{1}{2}})(1/\mathrm{diag}(\mathbf{\Sigma})^{\frac{1}{2}})^{\top}, \quad (1)$$

where $\mathrm{diag}(\cdot)$ denotes the diagonal elements of a matrix, $\mathbf{\Sigma} =$

$\frac{1}{n}(\mathbf{X}-\bar{\mathbf{X}})^\top(\mathbf{X}-\bar{\mathbf{X}})$ is a covariance matrix, n is the number of subjects, $\mathbf{X} \in \mathbb{R}^{t\times R}$ denotes the input time-series of a subject, t is the number of time points, R is the number of ROIs, and $\bar{\mathbf{X}} \in \mathbb{R}^{t\times R}$ denotes a matrix where each column is the mean of $\mathbf{X}$ over the time dimension. Tangent embedding $\vec{\mathbf{\Sigma}}$ can be obtained by mapping the covariance into tangent space via

$$\vec{\mathbf{\Sigma}} = \log_m(\mathbf{\Sigma}_G^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{\Sigma}_G^{-\frac{1}{2}}), \tag{2}$$

where $\log_m$ is matrix logarithm, $\mathbf{\Sigma}_G$ is the geometric mean (reference point) of covariance in a manifold, for example, the mean covariance matrix in Euclidean space: $\mathbf{\Sigma}_G = \frac{1}{n}\sum_{i=1}^{n}\mathbf{\Sigma}_i$. Here we propose replacing the input to tangent embedding $\mathbf{\Sigma}$ by a second-order correlation matrix $\breve{\mathbf{R}}$, where

$$\breve{\mathbf{R}} = \breve{\mathbf{\Sigma}}(1/\mathrm{diag}(\breve{\mathbf{\Sigma}})^{\frac{1}{2}})(1/\mathrm{diag}(\breve{\mathbf{\Sigma}})^{\frac{1}{2}})^\top, \tag{3}$$

and $\breve{\mathbf{\Sigma}} = \frac{1}{n}(\mathbf{R}-\bar{\mathbf{R}})^\top(\mathbf{R}-\bar{\mathbf{R}})$, then the tangent embedding of the second-order correlation can be obtained via

$$\vec{\mathbf{\Sigma}}_{\mathrm{TP}} = \log_m(\breve{\mathbf{R}}_G^{-\frac{1}{2}}\breve{\mathbf{R}}\breve{\mathbf{R}}_G^{-\frac{1}{2}}), \tag{4}$$

where the geometric reference point is $\breve{\mathbf{R}}_G = \frac{1}{n}\sum_{i=1}^{n}\breve{\mathbf{R}}_i$.

We name this proposed measure in Eq. (4) as the *Tangent Pearson embedding FC measure*, or simply *Tangent Pearson* (TP) because it can be seen as combination of the Pearson and tangent embedding FC measures with a *simple implementation*: replacing the covariance matrices used in the original tangent embedding method with the proposed correlation-based second-order FC matrices so that the deviation of each subject from the group is computed from the group mean second-order FC matrix.

The resulting FC matrices computed for each subject using these three FC measures are symmetric. Therefore, it is sufficient to keep only the upper/lower triangular parts. We chose to keep the upper triangular parts and also discarded the main diagonal FC values since they represented the self connectivity of an ROI with itself, which is redundant information. The remaining upper triangular values were flattened into a one-dimensional vector and used as FC features for each subject in all subsequent analyzes.

D. Statistical test of independence

Despite the aggregation of voxels' rs-fMRI time series and the estimation of FC features, acquisition site effects have been observed in the results of studies using these FC features. For example, Plitt et al. [24] identified large univariate differences in FC strength between three investigated sites (NYU, UCLA_1 and USM), showing the persistence of site effects into FC features. Parisot et al. [11] and Nielsen et al. [14] highlighted that significantly higher accuracies can be obtained on single-site studies in comparison to multi-site studies, however, this is also due to the reduced heterogeneity of both ASD and TC subjects in single-site studies. Before dealing with the dependence between ABIDE sites and FC features, we aim to first assess the statistical significance of this dependence. This assessment would provide a measure of the influence of site effects on FC features, and whether the effect is large enough to merit being accounted for during the modeling of ASD classifiers on ABIDE.

To conduct this statistical test, we employed the Hilbert-Schmidt independence criterion (HSIC), an empirical kernel-based statistical independence measure. It is superior to other kernel-based independence measures due to being simpler, converging faster and having a low sample bias w.r.t. the sample size [25], [26]. We firstly used the HSIC to measure the statistical dependence between sites and FC features. We then evaluated the statistical significance of such dependence using a hypothesis test derived for the HSIC [26].

1) Hilbert-Schmidt independence criterion: Given two multivariate random variables $\mathbf{X}$ and $\mathbf{Y}$ with associated probability distributions $P_{\mathbf{X},\mathbf{Y}}, P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$, the HSIC provides a non-parametric way of measuring their statistical dependence. It gives a measure of zero if they are independent, and a value greater than zero otherwise. The larger its value, the stronger the dependence between them. Empirically, given n realizations for the random variables $\mathbf{X} = \{\mathbf{x}_i\}$ and $\mathbf{Y} = \{\mathbf{y}_i\}$, the HSIC $\rho_h(\mathbf{X},\mathbf{Y})$ between $\mathbf{X}$ and $\mathbf{Y}$ is given by [25]

$$\rho_h(\mathbf{X},\mathbf{Y}) = \frac{1}{n^2}\,\mathrm{tr}(\mathbf{KHLH}), \tag{5}$$

where $\mathbf{K},\mathbf{H},\mathbf{L} \in \mathbb{R}^{n\times n}, \mathbf{K}_{i,j} = k_x(\mathbf{x}_i,\mathbf{x}_j)$ and $\mathbf{L}_{i,j} = k_y(\mathbf{y}_i,\mathbf{y}_j)$. $k_x(\cdot)$ and $k_y(\cdot)$ are two kernel functions, e.g. linear, polynomial, or radial basis function (RBF). $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is a centering matrix and tr$(\cdot)$ is the trace function.

We define $\mathbf{X}$ to be the random variable corresponding to FC features so that $\mathbf{x}_i$ contains a single subject's FC features. We define $\mathbf{Y}$ to be the random variable corresponding to the acquisition site, with $\mathbf{y}_i \in \mathbb{R}^{20}$ a one-hot encoding of each site (more detail in Sec. II-E). For the kernel functions, a linear kernel was used for $k_y(\cdot)$ due to the theoretical results in [27], where a correlation between HSIC and distribution divergence measure is guaranteed by linear kernel. RBF kernel was used for $k_x(\cdot)$ to model non-linear dependence between FC features and sites. We set the width parameter of the RBF kernel, σ, with the median distance between FC features.

2) Measure of significance: Gretton et al. [26] proposed to measure the statistical significance of an HSIC estimate based on a hypothesis test of independence for two random variables $\mathbf{X}$ and $\mathbf{Y}$ using the HSIC estimate $\rho_h(\mathbf{X},\mathbf{Y})$ as a test statistic. The test considers a null hypothesis of independence $\mathcal{H}_0 : P_{\mathbf{X},\mathbf{Y}} = P_{\mathbf{X}}P_{\mathbf{Y}}$ against an alternative hypothesis $\mathcal{H}_A : P_{\mathbf{X},\mathbf{Y}} \neq P_{\mathbf{X}}P_{\mathbf{Y}}$. Evidence for the acceptance of the null hypothesis is obtained by comparing the test statistic $\rho_h(\mathbf{X},\mathbf{Y})$ against a threshold T. If $\rho_h(\mathbf{X},\mathbf{Y}) \leq T$, the null hypothesis can be accepted. In other words, w.r.t. T, the HSIC estimate is sufficiently close to zero for independence between $\mathbf{X}$ and $\mathbf{Y}$ to be accepted. In [26], this threshold is set to be the $1-\lambda$ ($\lambda \in [0,1]$) quantile of the null distribution for the test statistic, as approximated by a two-parameter Gamma distribution.

E. Step 3: Domain adaptation

To tackle the inter-site heterogeneity, we take a multi-source domain adaptation approach called maximum independence domain adaptation (MIDA) [13]. We hypothesize that there exists inter-site difference in the FC features of subjects from different sites. Thus, extracting new features that are site-independent can potentially improve classification

performance. MIDA obtains these site-independent features in an unsupervised manner utilizing the empirical HSIC in Eq. (5) as a measure of dependence. Given a multivariate random variable $\mathbf{S}$ for the acquisition site, a projection map parameterized by $\mathbf{W}$, $\phi_{\mathbf{W}}(\cdot)$, and a random variable $\mathbf{X}$ for the subject FC features, MIDA aims to learn $\mathbf{W}$ so that the empirical HSIC between the random variables $\phi_{\mathbf{W}}(\mathbf{X})$ and $\mathbf{S}$ is close to zero. Thus, in the space $\phi_{\mathbf{W}}(\mathbf{X})$, the data are independent of their respective acquisition sites. Specifically, let $\mathbf{x}_i \in \mathbb{R}^k$ be the FC features of subject i, where $i \in \{1, \ldots, n\}$, k and n are the dimension of FC features and total number of samples, respectively. MIDA learns a low-dimensional feature $\mathbf{z}_i \in \mathbb{R}^h$, where $h \leq n$.

To use MIDA, we need to construct a site feature vector $\mathbf{d}_i \in \mathbb{R}^v$ for each subject to encode information about their respective site. The simplest is the site label information (e.g. CMU, KKI), which can be encoded with a one-hot scheme:

$$\mathbf{d}_{i,p} = \begin{cases} 1 \text{ if subject } i \text{ is from site } p \\ 0 \text{ otherwise,} \end{cases} \tag{6}$$

where $\mathbf{d}_{ip}$ is the pth entry of $\mathbf{d}_i$ and $v = 20$ (the number of sites). We concatenated site and FC feature vectors to augment subject features with site information: $\mathbf{x}_i = [\mathbf{x}_i, \mathbf{d}_i]$. Then we learned a mapping $\mathbf{W} \in \mathbb{R}^{n \times h}$ to project the augmented features to a new subspace where the new features for all subjects are minimally dependent on the site:

$$\mathbf{Z} = \mathbf{W}^\top \mathbf{K}, \tag{7}$$

where $\mathbf{K} \in \mathbb{R}^{n \times n}$, $\mathbf{K}_{i,j} = k_x(\mathbf{x}_i, \mathbf{x}_j)$ and each column of $\mathbf{Z} \in \mathbb{R}^{h \times n}$, $\mathbf{z}_i$, is the new feature representation of the augmented features. Next, we formulated the objective function by maximizing the preserved data variance while minimizing the statistical dependence on the site, i.e.

$$\max_{\mathbf{W}} -\operatorname{tr}\left(\mathbf{W}^\mathrm{T}\mathbf{KHLHKW}\right) + \mu \operatorname{tr}\left(\mathbf{W}^\mathrm{T}\mathbf{KHKW}\right), \tag{8}$$

where $\mu > 0$ is a hyper-parameter governing the emphasis of variance preservation against the level of independence achieved between the projected features $\mathbf{Z}$ and the site features. The solution can be found by forming $\mathbf{W}$ from the eigenvectors of the matrix $\mathbf{K}\left(-\mathbf{HLH} + \mu\mathbf{H}\right)\mathbf{K}$ corresponding to the h largest eigenvalues [13].

The hyper-parameters are μ, h and the kernel functions $k_x(\cdot)$, $k_d(\cdot)$. We used a linear kernel for $k_d(\cdot)$ and the RBF kernel for $k_x(\cdot)$ with the width parameter set to the median distance between FC features as in Sec. II-D.1 and optimized μ and h via a grid-search scheme detailed in Sec. II-H.

Theoretical analysis: According to the theoretical results in [27], for a classifier f trained on source domain samples, the upper bound of its generalization risk on target samples is:

$$\mathrm{R}_t(f) \leq \hat{\mathrm{R}}_s(f) + \rho_h(\mathbf{X}, \mathbf{D}) + \mathrm{O}\left(\sqrt{\frac{1}{n_s}\ln|\mathcal{H}|}\right) + \lambda^*, \tag{9}$$

where $\mathrm{R}_t(f)$, $\hat{\mathrm{R}}_s(f)$ are the (empirical) risk on target and source samples, respectively, $\rho_h(\mathbf{X}, \mathbf{D})$ denotes the HSIC between input FC matrix $\mathbf{X}$ and corresponding site feature matrix $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots \mathbf{d}_n]$, $\mathrm{O}(\cdot)$ denotes computational complexity, n_s is the number of labeled source domain samples, $|\mathcal{H}|$ denotes the complexity of hypothesis space, i.e., a metric to measure the set of all possible classification solutions ($|\mathcal{H}|$ equals to the total number of solutions if the solution set is finite), and $\lambda^* = \mathrm{R}_t(f^*) + \mathrm{R}_s(f^*)$ is the risk of an ideal hypothesis in theory, where $f^* = \arg\min_{f \in \mathcal{H}} \left(\mathrm{R}_s(f) + \mathrm{R}_t(f)\right)$. The last two terms can be viewed as constants in a classification task. Therefore, generalizbility can be improved via reducing the dependence between input FC data and site information, which is the optimization objective of MIDA. This theory also helps explain how the use of unlabeled target domain (test) data can improve learning performance.

F. Step 4: Incorporating phenotypic information

ABIDE has extensive phenotypic information. Including such features when training a classifier has been shown to be beneficial [28]. Several studies on autism have observed sex and age-related differences between ASD and TC. Werling & Geschwind [29] identified sex-differential genetic and hormonal factors that supported the observation that females are typically less frequently affected by ASD than males. In [30], age-matched ASD and TC children were found to have differences in FC. In fact, between ASD patients, FC differences have also been observed with respect to age [31], [32]. However, these findings are not yet conclusive [33].

Recent studies [8], [11] on ABIDE improved ASD classification accuracy by leveraging phenotypic information. We proceeded in a similar way to assess the impact of including phenotypes in ASD classifiers. We considered only sex, age, full IQ (FIQ), handedness and eye status at scan since the majority of subjects had such information present. For each categorical variable (handedness, sex and eye status at scan), a one-hot encoding scheme was used to construct phenotype features for each subject, which are *concatenated* with other features before feeding into a classifier. For subjects with missing values for FIQ and handedness, we used the same imputation method used in [8]. 1) *Handedness:* right hand dominance was assigned since most people are right-handed; 2) *FIQ:* the average IQ score of 100 was assigned.

G. Step 5: Classification

The impact of removing site effects can be assessed by comparing using the "raw" FC features (Sec. II-C) against site-independent features from MIDA as inputs to a classifier. Here we prefer linear (over deep) learning models to make isolating the impact of domain adaptation less complex and allow for a greater degree of interpretability [34], e.g. by visualizing the model coefficients to identify functional differences between ASD and TC. We chose three standard linear classifiers: ridge classifier (ridge regression with binary target values), logistic regression (LR), and support vector machine (SVM) from Scikit-learn [35]. For all models, the hyper-parameter values were selected via a grid-search scheme detailed in Sec. II-H.

H. Experimental setup

We designed the experiments with three objectives: 1) To test the statistical dependence between acquisition sites and

TABLE II
HYPER-PARAMETER SETTING FOR THE ML MODELS: h: THE NUMBER OF EIGENVECTORS IN MIDA; μ: THE WEIGHTING OF VARIANCE MAXIMIZATION IN MIDA; C: THE l_2 REGULARIZATION COEFFICIENT FOR THE THREE CLASSIFIERS ($\frac{1}{C}$ FOR LOGISTIC REGRESSION AND SVM).

ML Method	Hyper-parameter 1	Hyper-parameter 2
MIDA	h = 50, 150, 300	μ = 0.5, 0.75, 1.0
Ridge classifier	C = 0.25, 0.5, 0.75	N.A.
Logistic regression	C = 1, 5, 10	N.A.
SVM	C = 1, 5, 10	N.A.

FC features. 2) To assess the impact of the proposed second-order FC measure and site-dependence minimization on autism classification. 3) To extract biomarkers from the trained ML models for interpretation.

1) Algorithm setting.: We tested various models that can be constructed from the pipeline in Fig. 1, including both existing and proposed ones. We used CC200 as the default atlas. Table II lists the hyper-parameters for grid search. We evaluated all possible combinations of three values for each on the training data via five-fold cross validation (CV) to find the best setting. Specifically, the training data was further divided into five folds, where four folds were used for training and one fold was held out for validation. By repeating five times, the combinations of hyper-parameters with the highest averaged accuracy over the five folds were selected for final training.

2) Statistical test of independence.: We assessed the independence between sites and FC features using the statistical test in Sec. II-D. We set the significance level $\lambda := 0.05$ so that the probability of rejecting the null hypothesis when it is true is 0.01. In particular, given an observed HSIC estimate from the data, $\rho_h(\mathbf{X}, \mathbf{Y})$, the null hypothesis was set to be rejected at the 5% level if $\rho_h(\mathbf{X}, \mathbf{Y}) > t_{1-\lambda}$ with $t_{1-\lambda}$ being the 95% quantile of the estimated Gamma distribution. The kernel functions $k_y(\cdot)$ and $k_x(\cdot)$ were defined according to the empirical HSIC estimate detailed in Sec. II-D.1.

3) Prediction and comparison: We followed the terminology in [7] to consider the *intra-site* and *inter-site* prediction.

Intra-site prediction. This is the most commonly used setting [7], [9], [11], where the data from all 20 sites are mixed to form training/test sets with the same proportion of ASD/TC for stratified 10-fold CV. We compare MIDA-based models with those without using MIDA as baseline *raw* models to assess the impact of MIDA. We report the average accuracy and Area Under the Receiver Operating Characteristics (AUROC) over the 10 folds for each model. We also studied the impact of adding phenotypic features as in [8], [11]. For MIDA-based models, 'eye status at scan' was excluded as a phenotypic measure since it represents a site-specific protocol. Additionally, we evaluated impact of brain atlas by validating MIDA-based models on the CC200, CC400, and Harvard-Oxford (HO) atlases.

Inter-site prediction. This setting uses data from one individual site as testing data while training on the data from all 19 remaining sites to study the generalization performance to sites unseen in training. This is more challenging than the intra-site setting. The average accuracy/AUROC over the 20 sites will be reported. However, since each site has a different sample size, we computed the average by weighting the contribution of each site by sample size. Specifically, we measured the average accuracy by simply counting the total number of correct predictions across all sites from 20 runs and dividing by the total sample size. For AUROC, we similarly computed an overall measure across all sites. We also assessed the influence of phenotypic information as above.

Comparison with other studies. We compared our proposed method in both intra-site and inter-site settings with those in recent studies [7], [9], [11], and [12] (intra-site only). For inter-site setting, apart from reporting unweighted average accuracy only like in [7], [9], we will also report the results weighted by site sample sizes for fair comparison. Moreover, we studied two sample sizes, 871 and 1035, which have been studied in previous studies. For completeness, we also applied [11] to the larger sample size of 1035 using their implementation (available at https://github.com/parisots/population-gcn). To account for different stratifications in 10-fold CV between our study and [11], we used *the same stratification* as in [11] but also evaluated both methods using 5×10-fold CV to reduce the influence of the stratification. Furthermore, we conducted two-sample Welch t-tests to assess the statistical significance of performance improvements in intra-site evaluation and will report the corresponding p-values. The t-tests is not applicable to the inter-site setting, since we compute summary metrics without averaging and thus only have single representative values. We also conducted experiments with a recent domain adaptation approach proposed in [19] for multi-site fMRI data. However, its performance is not as good as MIDA, but with much higher computational cost. Therefore we only report the results of domain adaptation obtained by MIDA.

4) Biomarker extraction.: Linear classifiers evaluate $\mathbf{w}^\top\mathbf{x}$ to make a decision, where the model weight $\mathbf{w}$ is the learned decision hyperplane and $\mathbf{x}$ is the feature vector. In our context, the coefficients of $\mathbf{x}$ in $\mathbf{w}$ indicate the informativeness of each feature in $\mathbf{x}$ in classifying a subject as ASD or TC. If $\mathbf{x}$ is the (flattened) FC features, the values of the parameters in $\mathbf{w}$ will indicate which pairs of ROIs are important to distinguish ASD and TC. Positive and negative elements of $\mathbf{w}$ indicate ROI-ROI connections that are informative for ASD and TC, respectively. The larger the absolute value of the coefficient of a given ROI-ROI FC, the more informative this FC. We analyzed all the ten $\mathbf{w}$s from the 10-fold CV in the intra-site setting to extract ROI-ROI connections that are consistently the most informative for classification. Specifically, we identified the top 50 weights in absolute value for each fold and then ranked ROI-ROI FC connections for consistent occurrence in the top 50 features across all 10 folds, so that the most informative connections would be present in the top 50 features across all 10 folds. Where two ROIs had the same number of occurrence across folds (e.g. 7 times), we sorted the ranking by the absolute value of their respective coefficient averaged across folds.

III. RESULTS

A. Statistical test of independence

Table III shows the statistical test of independence between FC features and sites. For all FC measures, the null hypothesis

TABLE III

THE STATISTICAL TEST OF INDEPENDENCE: THE THRESHOLD VALUE GIVES THE 95% QUANTILE OF THE ESTIMATED GAMMA DISTRIBUTION. THE SAMPLE ESTIMATE GIVES THE SAMPLE HSIC ESTIMATE, WITH THE CORRESPONDING p-VALUE COMPUTED. THE DIFFERENCE IN THE THRESHOLD VALUES ARISES FROM THE DIFFERENT GAMMA DISTRIBUTION APPROXIMATIONS FOR DIFFERENT FC MEASURES.

FC measure	Threshold	Sample estimate	p-value
Pearson correlation	0.66	2.10	$< 10^{-5}$
Tangent	0.59	1.31	$< 10^{-5}$
Tangent Pearson (proposed)	0.59	1.21	$< 10^{-5}$

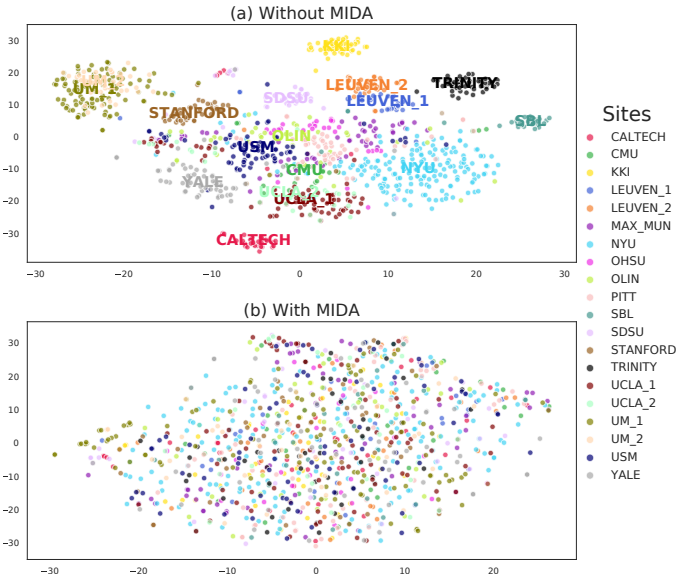

Fig. 2. The effect of domain-independent adaptation. A 2-D t-SNE projection of CC200-generated tangent Pearson FC features with (a) no site adaptation, (b) site adaptation by MIDA, using the scikit-learn t-SNE implementation with perplexity 30 and learning rate 10. In (a), key identifiable clusters are labeled while the other site labels (OHSU, MAX_MUN, and PITT) are omitted since their clusters are not well defined. In (b), site labels are omitted because MIDA removed the association between features and sites.

of independence between FC features and sites can be rejected with 95% confidence because the sample HSIC estimate exceeds the threshold. Also, the p-values are extremely small, giving greater evidence for rejecting the null hypothesis.

Next, we show the effect of domain-independent adapation by visualizing features in a 2-D space. We applied Principal Component Analysis to reduce the dimensionality of FC features to 50, and then employed t-distributed Stochastic Neighbour Embedding (t-SNE) [36] to project them to a 2-D space. Figure 2 shows the t-SNE projections of the proposed tangent Pearson FC features w.r.t. acquisition sites for both with/without site adaptation. In subplot (a) without adaptation, site-specific clusters can be identified (SDSU, SBL, etc.) while in subplot (b) with adaptation, there is a reduction of association between FC features and acquisition site. This further illustrates the site specificity of the ABIDE data.

B. Intra-site prediction

Figure 3 shows the intra-site prediction results. We analyze them with focus on the impact of (a) FC features, (b) MIDA,

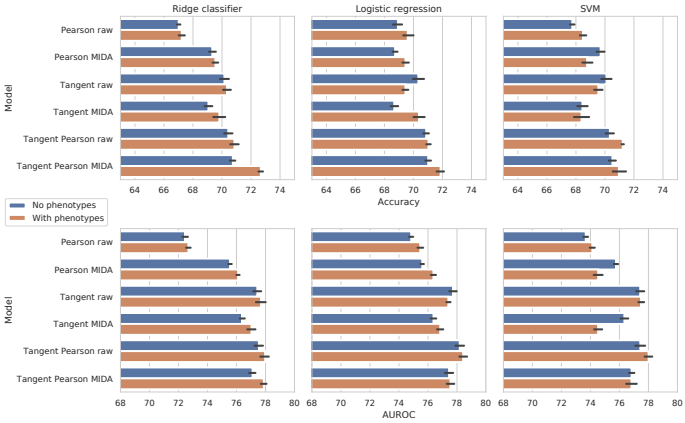

Fig. 3. Intra-site prediction with 5×10-fold CV. A comparison of the effect of site adaptation (raw baseline vs MIDA), classifier (ridge, logistic, SVM), FC measure (Pearson, tangent embedding, proposed tangent Pearson), and the inclusion of phenotypic information, on the average accuracy and AUROC.

and (c) phenotypic features, on the classification performance.

(a) Impact of FC measures. For the baseline models (raw, without MIDA/phenotypes), the three FC measures' performance is stable across classifiers and CV splitting settings. For each classifier, the highest accuracy was consistently obtained by the proposed tangent Pearson measure. In contrast, the Pearson correlation gave the lowest accuracy for all classifiers.

(b) Impact of MIDA. The effectiveness of MIDA seems to be sensitive to FC measure. Applying MIDA to Pearson features led to improvement in both accuracy and AUROC. In contrast, applying MIDA to tangent features led to worse results. Applying MIDA to TP embedding led to better accuracy (top of Fig. 3) but worse AUROC (bottom of Fig. 3) across different classifiers. A possible explanation for the performance drop is that there could be some loss of ASD/TC-specific information when learning site-independent features using the tangent FC with MIDA, so that some subjects are not properly represented for classification. In Sec. IV, we will discuss further studies that can potentially reduce such unwanted effects.

(c) Impact of phenotypic features. In most cases, adding the phenotypic features has improved the performance, which is consistent with the findings in [8], [11]. The best accuracy is 72.7%, obtained using TP+MIDA with phenotypes.

C. Impact of brain atlas

Table IV compares the intra-site results of MIDA with tangent Pearson FC and phenotypic features (TP MIDA) on three brain atlases: CC200, CC400, and HO. On the whole, there is no significant difference, except HO giving a relative lower accuracy and AUROC for Ridge classifier. Thus, following [8], [9], [11], we chose CC200 for our MIDA-based models.

D. Inter-site prediction

Figure 4 shows the inter-site results and we perform similar analyses as in the intra-site setting.

(a) Impact of FC measures. As in the intra-site setting, the tangent-based models have outperformed the Pearson

TABLE IV

IMPACT OF BRAIN ATLAS: RESULTS OF MIDA WITH TANGENT PEARSON AND PHENOTYPIC FEATURES USING THREE BRAIN ATLASES. THE 5×10-FOLD CV STANDARD DEVIATIONS (SD) OVER 5 DIFFERENT RANDOM SEEDS ARE IN PARENTHESES. ACC: ACCURACY. THE BEST RESULTS ARE IN **BOLD**, AND THE SECOND BEST ARE UNDERLINED.

	Ridge classifier		Logistic regression		SVM	
Atlas	ACC (SD)	AUROC (SD)	ACC (SD)	AUROC (SD)	ACC (SD)	AUROC (SD)
CC200	**72.7 (0.3)**	77.9 (0.4)	**71.9 (0.6)**	77.6 (0.5)	71.0 (1.0)	76.8 (0.8)
CC400	72.2 (0.3)	**78.3 (0.5)**	71.4 (0.4)	**78.0 (0.6)**	**71.1 (0.9)**	76.9 (1.3)
HO	71.5 (1.0)	77.0 (0.6)	71.6 (1.0)	77.1 (0.4)	71.0 (0.5)	**77.5 (0.5)**

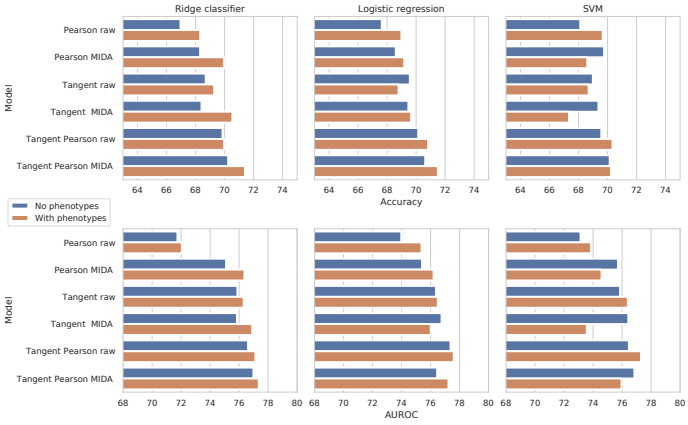

Fig. 4. Inter-site prediction (20 runs). As in Fig. 3, a comparison of the effect of site adaptation, classifier, FC measure, and phenotypic information on the weighted leave-one-site-out CV. Error bars are not shown here because each run uses data from one site as the test data so the variations across 20 runs/sites should not be interpreted as typical variations across multiple repetitions (e.g. in k-fold CV).

correlation-based ones. For the ridge and logistic regression classification models, the tangent Pearson measure achieves the highest accuracies of 69.9% (AUROC: 76.6%) and 70.1% (AUROC: 77.4%) respectively. The tangent measure achieves the highest accuracy of 69.6% (AUROC: 76.4%) with logistic regression. The Pearson correlation measure achieves its highest accuracy of 68.1% (AUROC: 73.1%) with SVM.

(b) Impact of MIDA. MIDA has an overall positive influence, with a reduced performance only for the tangent FC. For Pearson correlation and tangent Pearson measures, MIDA has increased the accuracy by 0.63% (AUROC: 0.90%) over baseline equivalents, averaged across the three classifiers. Tangent Pearson achieves the highest accuracy of 70.6% (AUROC: 76.4%) with MIDA and logistic regression, while its non-MIDA equivalent achieves an accuracy of 70.1% (AUROC: 76.4%). Applying MIDA to tangent FC features leads to no significant difference in accuracy/AUROC w.r.t. baseline models across all classifiers.

(c) Impact of phenotypic features. Adding phenotypic features improved the accuracy/AUROC of the baseline and MIDA models in most cases, e.g. an increases in accuracy by 1.05% (AUROC: 0.54%) on average over the three classifiers w.r.t. no phenotypic features. Overall, the highest inter-site accuracy of 71.4% (AUROC: 77.4%) is obtained with the phenotypic features, tangent Pearson FC, MIDA, and ridge classifier. Its non-phenotype equivalent has a lower accuracy

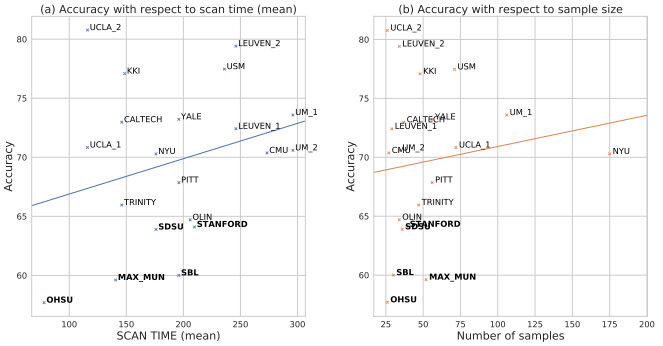

Fig. 5. Potential factors for site accuracy variation. The sites are visualized w.r.t. the accuracy obtained by the tangent Pearson MIDA ridge classifier without phenotypic features.

of 70.2% (AUROC: 77.0%).

(d) Factors for site accuracy variation. In inter-site prediction, the individual classification performance on each site varies a lot. Here, we investigate two factors that may affect the performance on an individual site:

1) *Mean length of rs-fMRI scan time*. We expect that having a longer experimental scan time increases the ability to detect differences between ASD and TC subjects.
2) *Number of samples collected*. Sites with small sample sizes may be under-represented in ABIDE and have distributions significantly different from other sites.

Figure 5 studies the correlations between site accuracy and site scan time or sample size from the results obtained by the model with tangent Pearson, MIDA, and ridge classifier but without phenotypic features. For scan time, a slight positive correlation between the mean scan time and the site accuracy can be identified (the left panel). Interestingly, the lowest scoring site (OHSU) has the lowest mean scan time of 78 seconds while all other sites exceed 116 seconds. Thus, it may be difficult to capture the ASD/TC differences effectively in such a short duration. Longer scan time can help remove noise from rs-fMRI and better capture differentiating signals. For site sample size, less conclusive relationship can be found with the observed site accuracy (the right panel).

E. Comparison with other studies

Intra-site comparison. Table V reports the intra-site results. In 10-fold CV, our TP MIDA ridge (with phenotypic features) on a sample size of 1035 outperforms the SOTA methods in both accuracy and AUROC with scores of 73.0% and 78.0%, respectively. This is an increase of 2.6% ($p = 0.09$) and 4.8% ($p < 10^{-2}$) in accuracy over to the SOTA [11] on 871 and 1035 subjects respectively. In 5×10-fold CV, TP MIDA ridge achieves an increase of 4.8% ($p < 10^{-2}$) and 4.1% ($p < 10^{-2}$) in accuracy over [11] on 871 and 1035 subjects respectively.

When including more training samples (all 1035 without QC), TP MIDA obtained a substantial improvement while the change is small for non-adaptation models, i.e. TP raw and [11]. The inter-site results in Table VI has similar observations. Section IV-C will discuss the relationship between those poor quality samples and domain adaptation effectiveness.

TABLE V

INTRA-SITE COMPARISON TO OTHER STUDIES. THE SAMPLE SIZE IS 871 IF QUALITY CONTROL (QC) WAS PERFORMED (DENOTED BY ✓), AND 1035 OTHERWISE (DENOTED BY ✗). WE FIRSTLY REPORT THE RESULTS OBTAINED USING THE SAME SPLIT SETTING OF 10-FOLD CV IN [11]. THEN WE SHOW THE RESULTS OBTAINED UNDER 5×10-FOLD CV WITH FOUR MORE CV SPLITTINGS GENERATED BY USING DIFFERENT RANDOM SEEDS. STANDARD DEVIATIONS FOR 10-FOLD CV WERE COMPUTED OVER 10 DIFFERENT PARTITIONS AND THOSE FOR 5×10-FOLD CV WERE COMPUTED OVER FIVE DIFFERENT RANDOM SEEDS. RESULTS OF [7], [9] ARE CITED FROM THE ORIGINAL PAPERS FOR REFERENCE, WHERE '-' INDICATES METRICS NOT AVAILABLE. FOR BOTH DATA WITH / WITHOUT QUALITY CONTROL. THE BEST RESULTS ARE IN **BOLD**, AND THE SECOND BEST ARE UNDERLINED.

		10-fold CV		5×10-fold CV	
Model	QC	ACC (SD)	AUROC (SD)	ACC (SD)	AUROC (SD)
TP MIDA	✗	**73.0 (3.9)**	78.0 (4.7)	**72.7 (0.3)**	77.9 (0.4)
TP raw	✗	71.6 (2.7)	**78.2 (3.6)**	70.9 (0.6)	**78.0 (0.7)**
Phenotype only	✗	57.8 (4.7)	60.3 (4.5)	57.8 (0.7)	59.9 (0.3)
Almuqhim [12]	✗	70.8 (-)	- (-)	- (-)	- (-)
Parisot [11]	✗	68.2 (3.7)	75.2 (3.8)	68.6 (0.3)	75.2 (0.5)
Heinsfeld [9]	✗	70.0 (-)	- (-)	- (-)	- (-)
TP MIDA	✓	70.0 (7.6)	**75.5 (7.3)**	69.7 (0.5)	75.9 (0.5)
TP raw	✓	69.2 (7.5)	75.1 (7.1)	**70.1 (0.8)**	**76.5 (0.9)**
Phenotype only	✓	58.6 (5.0)	69.4 (5.1)	59.4 (0.6)	60.6 (2.9)
Parisot [11]	✓	**70.4 (3.9)**	75.0 (4.6)	67.9 (1.3)	73.3 (0.9)
Abraham [7]	✓	66.9 (2.7)	- (-)	- (-)	- (-)

In the 10-fold CV, our TP raw ridge (with phenotypic features, without domain adaptation) on 1035 subjects achieves 71.6% in accuracy and 78.2% in AUROC, outperforming neural network based models [9], [11] by at least 1.2% in accuracy and 3.2% in AUROC. It obtained statistically significant increases in accuracy (3.4%, $p = 0.02$) and AUROC (3%, $p = 0.05$) at the 10% level w.r.t. to [11] on 1035 subjects. In the 5×10-fold CV, we obtained p-values less than 1% when comparing the accuracy/AUROC of TP raw ridge on 1035 subjects with both 871/1035-subject variants of [11]. On 871 subjects, our TP raw LR achieves 70.1% in accuracy and 76.5% in AUROC, ourperforming [11] and even TP MIDA LR, showing the effectiveness of the proposed TP FC measure.

Inter-site comparison. Table VI shows the inter-site performance. We firstly observe that the weighted site accuracy and AUROC scores are mostly higher than the unweighted results. This is expected because from the right panel of Fig. 5, sites with lower accuracy tends to have a smaller sample size. Secondly, our TP MIDA ridge (with phenotypic features) on 1035 samples achieves the highest (weighted) accuracy of 71.4%, improving the model [11] on 871 and 1035 subjects by 2.9% and 2.8% in accuracy respectively. Without domain adaptation, our TP raw ridge (with phenotypic features) also achieves a higher (weighted) accuracy (and AUROC) than [11] when the sample size is 1035, an increase in accuracy by 1.5% and 1.4% w.r.t. [11] on 871 and 1035 subject respectively.

Across both sample sizes (871/1035), our proposed domain adaption and baseline models improve over the models in [7], [9] in unweighted accuracy. On 1035 subjects, our TP MIDA ridge model has the highest accuracy of 70.5%, improving over [9] and [7] by 5.5% and 3.7%, and our TP raw ridge model improves over [9] and [7] by 3.5% and 1.7%, respectively.

TABLE VI

INTER-SITE COMPARISON, I.E. LEAVE ONE SITE OUT CV (LOSOCV). THE NOTATIONS IN THIS TABLE ARE THE SAME AS IN TABLE V.

		Unweighted		Weighted	
Model	QC	ACC (SD)	AUROC (SD)	ACC	AUROC
TP MIDA	✗	**70.5 (7.6)**	**76.7 (8.8)**	**71.4**	**77.4**
TP raw	✗	68.5 (9.6)	76.3 (8.4)	70.0	77.1
Phenotype only	✗	56.4 (9.8)	57.3 (12.4)	57.9	59.6
Parisot [11]	✗	68.3 (5.6)	75.3 (7.3)	68.6	75.7
Heinsfeld [9]	✗	65.0 (1.4)	- (-)	-	-
TP MIDA	✓	**68.4 (8.3)**	75.6 (9.2)	69.3	75.3
TP raw	✓	68.1 (8.9)	**75.9 (9.1)**	**70.3**	**75.7**
Phenotype only	✓	59.2 (11.2)	58.1 (15.9)	59.1	58.8
Parisot [11]	✓	68.4 (6.3)	73.7 (7.0)	68.5	73.7
Abraham [7]	✓	66.8 (5.4)	- (-)	-	-

F. Extracting biomarkers

For the proposed FC measure, tangent Pearson, we study the respective ROI-ROI connections that have the most significant influence on the classification performance. These influential ROI-ROI connectivities can act as neurological biomarkers for researchers to investigate and further understand the difference in brain connectivity between ASD and TC.

To extract these biomarkers, we used the CC200 atlas to firstly define ROIs and the weights from the TP raw LR model (without phenotypic features) to indicate which ROI-ROI connections are most important for ASD/TC classification. The logistic regression classifier was used because it achieved the highest accuracy and AUROC for the tangent Pearson FC in the intra-site setting without phenotypic features. Phenotypic features were omitted because only ROI-ROI connections were of interest. The top five most positive and negative weights, for ASD and TC respectively, were extracted as in Sec. II-H.

The CC200 atlas is derived from the clustering of individual voxel BOLD time courses so the resulting atlas has no well-defined labels for the ROIs. To generate labels, we used the centre of mass for each ROI to locate the closest matching ROIs from the Harvard-Oxford brain atlas as a point of reference. Where no label could be found for a given CC200 ROI, we set its label to "None".

Figure 6 shows the top 10 most important ROI-ROI connections for classifying subjects as ASD or TC for illustration. Red and blue connections give the top five most important ROI-ROI connections for classifying subjects as ASD and TC, respectively. Increasing values of the red connections will bias the model decision towards ASD, while increasing values of the blue connections will bias the model decision towards control (non-ASD). Note that we omitted 180 of 200 ROIs from the figure to clearly show these connections.

IV. DISCUSSION

A. Intra-site and inter-site evaluation

For intra-site setting, k-fold CV is widely used in many studies. However, different random partitions can lead to significant variation of results, as seen by comparing the 10-fold CV and 5×10-fold CV results. Here we *recommend* $n \times k$-fold CV, where the effect of random split settings is reduced, and the model evaluation is more stable.

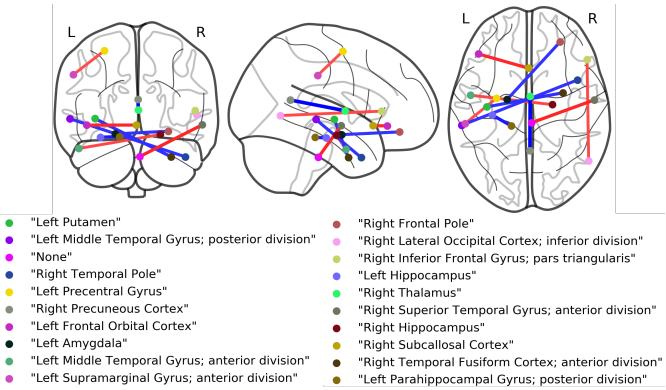

Fig. 6. Biomarker visualization. Extracted biomarkers using Python package Nilearn [37]. From left to right: the frontal, axial and lateral views of the brain are visualized. 'L' and 'R' correspond to the left and right hemisphere respectively.

For inter-site setting, due to significant sample size differences across sites, we recommend the weighted average accuracy, which represents $\frac{\text{total correct predictions}}{\text{total samples}}$ but the unweighted version does not. In the rest of this section, our discussion will be based on the 5×10-CV and weighted leave-one-site-out CV results for intra-site and inter-site settings, respectively.

B. Effectiveness of site-dependence minimization

We studied the impact of minimizing dependence between acquisition site and subject FC features. In most cases, removing the dependence between site and FC features led to an improved performance in both intra-site and inter-site settings (Figs. 3 and 4). MIDA-based models using the proposed tangent Pearson FC outperformed recent SOTA approaches [9], [11], achieving new SOTA performance of 72.7% (73% in the CV setting of [11]) in intra-site accuracy (AUROC: 77.9%), and 71.4% in inter-site accuracy (AUROC: 77.4%), corresponding to increases of 4.8% and 2.9% (AUROC: 4.6% and 3.7%) w.r.t. [11], respectively. These results highlight the value of minimizing the dependence between FC features and acquisition sites for improving autism classification.

C. Low-quality samples in domain adaptation

For both intra-site and inter-site settings, we observed accuracy improvement by MIDA-based models when including more *low-quality* samples in training. In contrast, including them has much less effect on the classification accuracy of other models. On one hand, these samples may be helpful in estimating the site data distribution so that MIDA can extract better domain-independent features. On the other hand, their phenotypic features are not necessarily low quality and can also contribute to the autism prediction.

D. Effectiveness of second-order functional connectivity

The proposed tangent Pearson embedding FC measure combines two existing FC measures: Pearson correlation and tangent embedding. Without MIDA, we observed that this new second-order FC measure can outperform previous SOTA methods when supplemented with phenotypic features and a linear classifier. We achieved the highest accuracy (among raw) of 70.9% (AUROC: 78.0%) in the intra-site experiment (Fig. 3), and an accuracy of 70.0% (AUROC: 77.1%) in the inter-site setting (Fig. 4), improving upon [11] by 3.0% and 1.6% in accuracy (4.7% and 3.4% in AUROC), respectively. The proposed FC measure becomes more attractive if considering that those SOTA methods employ complex neural networks taking a long time to train (e.g. *32 hours* for [9]). The *linear* classifiers investigated in this study can achieve improved results with only *several minutes* of training when leveraging the proposed TP FC features and phenotypic information.

E. Robustness of biomarkers

To study the robustness of biomarkers (model weights), we computed the Pearson correlation and the number of overlapped top weights between the model trained on full ABIDE data (1035 subjects) and each of the 20 models trained on leave-one-site-out data. All models were trained using TP + MIDA (linear kernel) + Logistic Regression with fixed hyperparameters. All classifier coefficients were transformed to the original (20,100 dimensional) feature space for comparison. The average Pearson correlation is 0.904 ± 0.042, and the average number of overlapped top weights is 34.6 ± 4.8 out of 50. Thus adding one site will not change the learned biomarkers dramatically, with 69% consistent on average.

F. Limitations and future directions

Two limitations are apparent when using MIDA to minimize site dependence. Firstly, in a few cases, particularly with the tangent FC, MIDA led to a degradation in intra/inter-site accuracy and AUROC w.r.t. baseline models. A potential cause is the difficulty in preserving relationships between projected FC features and target labels (ASD/TC). Though the variance in the original FC features can be preserved with MIDA, the alignment between projected features and target labels may not be (fully) preserved. This is particularly important for FC features derived from rs-fMRI for autism classification since the underlying signal defining autism is not well marked due to the heterogeneity of ASD. Therefore, developing methods that can use the training data labels to align subject FC features without overfitting may help unlock more potential from ABIDE in multi-site autism classification.

Secondly, since MIDA is a transductive learning method, adding new subjects to the experimental dataset would require a new domain-independent subspace to be learned to account for the new data before predictions can be made. However, such retraining is of low cost because our learning approach is time efficient. For example, we trained the 21 models in Sec. IV-E with CPU only (Intel i7-12700K), and the average time consumed for the whole training process is less than six minutes. Another research focus is developing an inductive domain adaptation approach that alleviate this problem.

The proposed second-order FC measure and MIDA approach affect only the feature representation of each subject to improve multi-site autism classification. The obtained performance is mostly similar across three different classifiers (Figs. 3 and 4). Therefore, we hypothesize that to improve multi-site

autism classification, it is important to 1) design more powerful FC measures or other measures from raw fMRI data, and 2) directly target and remove site agglomerative effects.

V. Conclusion

This paper focused on improving multi-site autism classification. We proposed a new second-order functional connectivity measure called tangent Pearson embedding for more discriminative features, and took a site-dependence-minimization domain adaptation approach to tackle the heterogeneity in the multi-site ABIDE database. We confirmed the significance of this study via a statistical independence assessment between acquisition sites and FC features. The intra- and inter-site classification results show that models with the proposed FC measure, site-dependence minimization, and phenotypic features outperformed state-of-the-art methods.

References

[1] J. Baio, "Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010." *Morbidity and Mortality Weekly Report. Surveillance Summaries*, vol. 63, no. 2, pp. 1–21, 2014.

[2] B. A. Zielinski, M. B. Prigge, J. A. Nielsen, A. L. Froehlich *et al.*, "Longitudinal changes in cortical thickness in autism and typical development," *Brain*, vol. 137, no. 6, pp. 1799–1812, 2014.

[3] L. Zwaigenbaum and M. Penner, "Autism spectrum disorder: advances in diagnosis and evaluation," *British Medical Journal*, vol. 361, 2018.

[4] Y. Du, Z. Fu, and V. D. Calhoun, "Classification and prediction of brain disorders using functional connectivity: Promising but challenging," *Front. Neurosci.*, vol. 12, p. 525, 2018.

[5] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *NeuroImage*, vol. 145, pp. 137–165, 2017.

[6] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson *et al.*, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[7] A. Abraham, M. P. Milham, A. D. Martino, R. C. Craddock, D. Samaras, B. Thirion, and G. Varoquaux, "Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example," *NeuroImage*, vol. 147, pp. 736 – 745, 2017.

[8] N. C. Dvornek, P. Ventola, and J. S. Duncan, "Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks," in *ISBI*, 2018, pp. 725–728.

[9] 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.

[10] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, and D. Rueckert, "Metric learning with spectral graph convolutions on brain connectivity networks," *NeuroImage*, vol. 169, pp. 431–442, 2018.

[11] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, "Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease," *Medical Image Analysis*, vol. 48, pp. 117–130, 2018.

[12] F. Almuqhim and F. Saeed, "ASD-SAENet: a sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using fMRI data," *Front. Comput. Neurosci.*, vol. 15, p. 654315, 2021.

[13] K. Yan, L. Kou, and D. Zhang, "Learning domain-invariant subspace using domain features and independence maximization," *IEEE T. Cybern.*, vol. 48, no. 1, pp. 288–299, 2017.

[14] J. A. Nielsen, B. A. Zielinski, P. T. Fletcher, A. L. Alexander, N. Lange, E. D. Bigler, J. E. Lainhart, and J. S. Anderson, "Multisite functional connectivity MRI classification of autism: ABIDE results," *Front. Hum. Neurosci.*, vol. 7, p. 599, 2013.

[15] J. G. Castrillon, A. Ahmadi, N. Navab, and J. Richiardi, "Learning with multi-site fMRI graph data," in *Asilomar Conference on Signals, Systems and Computers*, 2014, pp. 608–612.

[16] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2009.

[17] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[18] E. Moradi, B. Khundrakpam, J. D. Lewis, A. C. Evans, and J. Tohka, "Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data," *NeuroImage*, vol. 144, pp. 128–141, 2017.

[19] M. Wang, D. Zhang, J. Huang, P.-T. Yap, D. Shen, and M. Liu, "Identifying autism spectrum disorder with multi-site fmri via low-rank domain adaptation," *IEEE Trans. Med. Imaging*, vol. 39, no. 3, pp. 644–655, 2020.

[20] L. Rabany, S. Brocke, V. D. Calhoun, B. Pittman, S. Corbera, B. E. Wexler, M. D. Bell, K. Pelphrey, G. D. Pearlson, and M. Assaf, "Dynamic functional connectivity in schizophrenia and autism spectrum disorder: Convergence, divergence and classification," *NeuroImage: Clinical*, vol. 24, p. 101966, 2019.

[21] G. Varoquaux, F. Baronnet, A. Kleinschmidt, P. Fillard, and B. Thirion, "Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling," in *MICCAI*, 2010, pp. 200–208.

[22] R. C. Craddock, G. A. James, P. E. Holtzheimer III, X. P. Hu, and H. S. Mayberg, "A whole brain fMRI atlas generated via spatially constrained spectral clustering," *Hum. Brain Mapp.*, vol. 33, no. 8, pp. 1914–1928, 2012.

[23] N. Makris, J. M. Goldstein, D. Kennedy, S. M. Hodge, V. S. Caviness, S. V. Faraone, M. T. Tsuang, and L. J. Seidman, "Decreased volume of left and total anterior insular lobule in schizophrenia," *Schizophrenia Research*, vol. 83, no. 2-3, pp. 155–171, 2006.

[24] M. Plitt, K. A. Barnes, and A. Martin, "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards," *NeuroImage: Clinical*, vol. 7, pp. 359–366, 2015.

[25] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," in *ALT*, 2005, pp. 63–77.

[26] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola, "A kernel statistical test of independence," in *NeurIPS*, 2008, pp. 585–592.

[27] S. Zhou, W. Li, C. R. Cox, and H. Lu, "Side information dependence as a regularizer for analyzing human brain conditions across cognitive experiments," in *AAAI*, 2020, pp. 6957–6964.

[28] M. N. Parikh, H. Li, and L. He, "Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data," *Front. Comput. Neurosci.*, vol. 13, p. 9, 2019.

[29] D. Werling and D. Geschwind, "Sex differences in autism spectrum disorders," *Current Opinion in Neurology*, vol. 26, pp. 146–53, 04 2013.

[30] R. K. Kana, J. O. Maximo, D. L. Williams, T. A. Keller, S. E. Schipul, V. L. Cherkassky, N. J. Minshew, and M. A. Just, "Aberrant functioning of the theory-of-mind network in children and adolescents with autism," *Molecular Autism*, vol. 6, no. 1, p. 59, 2015.

[31] L. Uddin, K. Supekar, and V. Menon, "Reconceptualizing functional brain connectivity in autism from a developmental perspective," *Front. Hum. Neurosci.*, vol. 7, p. 458, 08 2013.

[32] K. Supekar, L. Q. Uddin, A. Khouzam, J. Phillips, W. D. Gaillard, L. E. Kenworthy, B. E. Yerys, C. J. Vaidya, and V. Menon, "Brain hyperconnectivity in children with autism and its links to social deficits," *Cell Reports*, vol. 5, no. 3, pp. 738–747, 2013.

[33] R.-A. Müller, P. Shih, B. Keehn, J. R. Deyoe, K. M. Leyden, and D. K. Shukla, "Underconnected, but how? A survey of functional connectivity MRI studies in autism spectrum disorders," *Cerebral Cortex*, vol. 21, no. 10, pp. 2233–2243, 2011.

[34] C. Molnar, *Interpretable Machine Learning*, 2020.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[36] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[37] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *Front. Neuroinformatics*, vol. 8, p. 14, 2014.