Transfer Learning Approaches in Classification of Mental Disorders using Neuroimaging

by

Bonian Lu

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 7, 2022

Keywords: Functional MRI, Functional connectivity, autism spectrum disorders, transfer learning, domain adaptation, feature identification

Approved by

Gopikrishna Deshpande, Chair, Professor, Department of ECE, Auburn University
Thomas Denney, Professor, Department of ECE, Auburn University
Jeffrey Katz, Professor, Department of Psychological Sciences, Auburn University
Stanley Reeves, Professor, Department of ECE, Auburn University

Abstract

Resting-state functional magnetic resonance imaging (rs-fMRI) data is widely used to characterize brain function in health and disease. Specifically, brain networks obtained from rs-fMRI based functional connectivity (FC) have been used to investigate disconnection syndromes in psychiatry. A good example is Autism Spectrum Disorder (ASD), wherein statistical group differences between ASD and controls in terms of FC have been widely reported, with FC being predominantly weaker in ASD. The aberrant behaviors of ASD patients have been found to be associated with abnormal patterns of RSFC, including the reduced correlation between frontal and posterior brain networks and local FC strengthening with long-distance FC reduction. Statistical group comparison suffers from the disadvantage that it does not possess the ability to predict the outcome (such as diagnostic status) in a novel subject. Machine learning models have been used for individual subject-level characterization as an alternative.

Deep learning models outperform traditional machine learning methods in diagnostic classification. For ASD, open-source datasets such as ABIDE (Autism Brain Imaging Data Exchange) have accelerated the application of deep learning to the diagnostic category. However, overfitting is the main issue that constrains the validity and generalizability of deep learning and traditional machine learning. Overfitting refers as a well-trained deep neural network that can achieve high prediction accuracy in the training dataset but has poor prediction accuracy in the unseen test dataset. The primary cause of overfitting is the relatively small sample size in training dataset compared to the high dimensionality of the feature space, known as the “curse of dimensionality”. In addition, to address the sample size issue, data is being increasingly aggregated across sites to form large databases such as ABIDE, given that acquiring a large number of subjects from a single site can be costly and time-consuming. However, the

problem with such an approach is wide inter-site variability in data characteristics that are non-neural in origin: differences across MRI vendors, pulse sequences, sequence parameters, sampling of patient populations, and data preprocessing pipelines. Consequently, machine learning (ML) approaches, including deep learning (DL), tend to produce high accuracy in diagnostic classification using single-site data but poor accuracy using multi-sites data. This lack of generalizability in the model is the most significant barrier to adopting ML and DL in neuroimaging-based diagnostics. We propose two approaches that address this issue.

We used a VAE-CNN (variational autoencoder convolutional neural network) transfer learning model in the first project. Because it is harder to acquire and aggregate patient population data than healthy controls, comparatively larger samples are available from healthy controls in the public domain, we propose to address overfitting by using larger healthy samples to learn the neural signature of healthy controls, with the aim of "transferring" that learning into the context of discriminating clinical populations. Here, we investigate the utility of transfer learning from HCP (human connectome project) healthy control data for improving the classification of individuals with ASD from their healthy peers in ABIDE data. We identify the biomarkers contributing to classification using the Layer-Wise Relevance Propagation (LRP) algorithm. Results show that the proposed transfer learning method outperforms state-of-the-art deep learning methods in ASD classification, especially when training and testing data are drawn from different data acquisition sites.

In the second project, we propose domain adaptation for improving the generalizability of neuroimaging-based diagnostic classification. Domain adaptation aims to improve classification performance in a given target domain by utilizing the knowledge learned from a different source domain by making data distributions of the two domains as similar as possible. To validate the

utility of domain adaptation for classifying multi-site fMRI data, we developed a variational autoencoder – maximum mean discrepancy (VAE-MMD) model for three-way diagnostic classification of Autism, Asperger's syndrome, and controls. In domain adaptation, we chose ABIDEII (Autism Brain Imaging Data Exchange) as the target domain data and ABIDEI as the source domain data. The results show that the domain adaptation approach achieved superior test accuracy of ABIDEII compared to baseline methods using just ABIDEII for classification. In addition, we augmented the source domain with additional healthy control subjects from Healthy Brain Network (HBN) and Amsterdam Open MRI Collection (AOMIC) datasets, enabling transfer learning to improve classification performance. Finally, we compared domain adaptation and combined statistical ComBat harmonization in this study. The result demonstrated that the domain adaptation model could be improved when combined with statistical methods. We openly share our data and model so that the neuroimaging community can explore the possibility of further improvement of the model by utilizing the ever-increasing amount of healthy control fMRI data in the public domain.

Acknowledgements

First and foremost, I wish to express my sincere gratitude to my supervisor, Dr. Gopikrishna Deshpande, for the continuous support throughout my Ph.D. study and related research. His guidance has been of invaluable help in the research and writing of this thesis.

Apart from my advisor, I would like to thank the rest of my thesis committee members: Dr. Thomas Denney, Dr. Jefferey Katz, Dr. Stanley Reeves, and my university reader Dr. Sridhar Krishnamurti, for their insightful comments and encouragement.

My sincere gratitude also goes to my fellow labmates, Sinan Zhao, Yun Wang, and Ranga Deshpande, for the stimulating discussions, encouragement, and time working with me. Especially Ranga Deshpande put a lot of work into revising my papers, and these papers could not complete without his help.

In addition, many persons have supported the datasets used in this study for the data provider. Primary support for the work by Adriana Di Martino was provided by the (NIMH K23MH087770) and the Leon Levy Foundation. Primary support for the work by Michael P. Milham and the INDI team was provided by gifts from Joseph P. Healy and the Stavros Niarchos Foundation to the Child Mind Institute and an NIMH award to MPM (NIMH R03MH096321). Primary support for the work by Adriana Di Martino and her team was provided by the National Institute of Mental Health (NIMH 5R21MH107045). Primary support for the work by Michael P. Milham and his team was provided by the National Institute of Mental Health (NIMH 5R21MH107045); Nathan S. Kline Institute of Psychiatric Research). Additional support was provided by gifts from Joseph P. Healey, Phyllis Green, and Randolph Cowen to the Child Mind Institute.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AE	Autoencoder
AD	Alzheimer’s disease
ADNI	Alzheimer’s Disease Neuroimaging Initiative
ASD	Autism Spectrum Disorder
AUC	Area under curve
BOLD	Blood Oxygenation Level Dependent
CNN	Convolutional Neural Network
CT	Computed Tomography
DL	Deep learning
DMN	Default Mode Network
DPARSF	Data Processing Assistant for Resting State fMRI
dHB	Deoxygenated Hemoglobin
EC	Effective Connectivity
FC	Functional Connectivity
fMRI	Functional Magnetic Resonance Imaging
FWHM	full width half maximum
HCP	Human Connectome Project
HRF	Hemodynamic response function
ML	Machine learning
MLP	Multilayer perceptron
MNI	Montreal Neurological Institute
MRI	Magnetic Resonance Imaging
RS-fMRI	Resting state fMRI
ROC	Receiver operating characteristic
ROI	region of Interest
SNR	signal-to-noise ratio
SPM	Statistical Parametric Mapping
SSL	Semi-supervised learning
TL	Transfer learning
TR	Time of Repetition
VAE	Variational Autoenco

Chapter 1

Introduction

1.1 Functional Magnetic Resonance Imaging (fMRI)

Functional magnetic resonance imaging (fMRI) is a technique that relies upon the measurement of the magnetic spin-spin relaxation time T2*, which reflects the changes in local concentrations of paramagnetic deoxyhemoglobin (HbR) [1]. The blood oxygen level dependent (BOLD) signal is sensitive primarily to decreasing HbR, so it can be detected to produce a localized map of activity in the human brain [2]. Compared with other noninvasive assessments of brain function techniques such as electroencephalography (EEG) and magnetoencephalography (MEG), fMRI has advantages of better spatial resolution and localization in the millimeter range per voxel. In recent decades, fMRI has become a primary technique for studying the human brain in research and clinical applications [3] [4]. However, one study compared the BOLD signal to invasive EEG and MEG signals. It showed that the BOLD contrast signal mainly reflects the neuron's integrative processes within its body instead of the neuron's response by itself [5]. Therefore, a deconvolution method has proven necessary to extract a latent neural response from BOLD signals. More details about deconvolution will be illustrated in the deconvolution section next.

The two major categories of studies on fMRI data are task fMRI and resting-state fMRI. Resting-state fMRI(RS-fMRI) is applied to detect the baseline or spontaneously BOLD variance in the absence of any task. The basic theory of RS-fMRI is that spontaneous neural fluctuations can reveal some fundamental brain characteristics in both structural and functional aspects [6]. Thus RS-fMRI study plays a crucial role in neuroimaging research to analyze highly correlated subnetwork referred to as the default mode network (DMN) activated at rest. Some studies found

that a specific network pattern maintains housekeeping functions such as vascular perfusion, heart rate, or breathing [7] [8]. In clinical applications, spontaneous fMRI connectivity patterns have been used as a diagnostic and prediction tool in multiple psychiatric disorder areas such as autism, schizophrenia, and Alzheimer's disease (AD) [9] [10] [11], etc.

1.2 fMRI data preprocessing

The original fMRI data contains imperfections from head movement, spontaneous neuron activities, and intrinsic electron thermal noises. Proper preprocessing is necessary for further statistical analysis [12]. Based on various signal and image processing methods, the preprocessing of fMRI forms a pipeline of multiple steps. Many studies found that preprocessing poses a significant impact on the BOLD fMRI data [13] [14] [15]. However, among these studies, no consensus has been made on choosing steps performed in the preprocessing pipeline. Before data analysis in the following projects, this paper discussed several significant preprocessing steps, including motion correction, slice timing correction, and spatial normalization. Other standard measures include detrending, covariates regression, and bandpass filtering. In most studies, preprocessing is implemented by several toolboxes, including CONN, SPM for Matlab, and FSL for Mac OS X and Linux.

1.2.1 Slice-Timing correction

Slice-timing correction generally enhances the statistical analysis power in fMRI studies. One fMRI 3D image is constructed from a series of 2D slices. All slices from the image are acquired by separated times from Time of Repetition (TR). Thus, it makes all slices acquired at different times, or say different slices in a 3D image, are observed at varying time points. In further statistical analysis, we assume that all slices are obtained simultaneously. To correct the acquisition times of each slice, we adjust all voxel time series to a standard slice via sinc-

interpolation. Each slice has a different acquisition time, so each slice typically receives a specific interpolation.

1.2.2 Motion Correction

fMRI data analysis assumes that all time points in a time series represent a value from a single location. However, the subject will inevitably keep still in the scanner in a long-term experiment session. Instruction and training can reduce the head movement noise, but often not enough. In the task-related fMRI, head motion is correlated with task setting. Even in resting-state fMRI, head movement remains a problem [16] [17]. The shift from head movement can cause noise and uncertainty in fMRI analysis, so motion effects have to be removed or reduced in the data preprocessing pipeline.

Based on the theory assuming that head motion is a 3D rigid body movement, the motion can be parameterized by 6 degrees of freedom (DOF). 6 DOF contain three translation parameters and three rotation parameters. A reference image is set up first for each session, and other images are realigned to it by rigid body transformation parameters. Motion correction parameters can also be used in the nuisance regression step.

1.2.3 Spatial Normalization

To reduce the individual differences across subjects, spatial normalization can match the geometry of each subject into an “atlas” space. The goal of normalization is to find a nonlinear transformation that makes segmentation in an anatomical image, which can enhancing the statistical power in fMRI data analysis. There are three steps to normalized fMRI data. First, T1 (anatomical) and T2* (temporal) images are coregistered into the same space. Second, transforming anatomical images from different subjects into a known standard space, such as the

MNI152 template. Third, applying volume-based fMRI data transformation by using previously obtained parameters.

1.2.4 Spatial Smoothing

Spatial smoothing is a commonly preprocessing method to improve the signal-to-noise ratio (SNR) and inter-subject registration and further improve the ability of a statistical method to detect actual activations [18]. The opposite consequence of spatial smoothing is it will reduce the spatial resolution of the image. A typical way to implement spatial smoothing is weighting fMRI images with a width fixed Gaussian kernel. The width of the kernel decides how much the images will be smoothed. However, an increasing number of articles found that spatial smoothing has complex effects on brain networks [19] [20] [21]. The results include overemphasis of short distance path, distortion of network centrality measurement, ROI based functional network disconnection, etc. In this paper, no spatial smoothing was applied to avoid distortion in the functional network from the fMRI dataset.

Other than the steps discussed above, the 0.01~0.1 bandpass filtering and WM and CSF nuisance regression were also applied in the projects. We used several standard tools in preprocessing methods including CONN, SPM for Matlab, and FSL for Mac OS X and Linux. In this study, we used CONN and DAPRSF toolboxes on Matlab.

1.3 Blind Deconvolution

As mentioned above, the BOLD contrast signal is not directly measuring the human brain activity. Instead, even if the details of the relationship between the BOLD signal and the human brain activity are not fully understood, it can be treated as a filter version of brain activity combined with the hemodynamic response function (HRF) [22]. Assuming that the BOLD signal response to latent neural activity is linear, the BOLD signal can be modeled due to the

convolution between neural activity and HRF. This theory raises some issues in brain function analysis, especially in Granger causality measures of effective connectivity [23] [24] [25]. A deconvolution method is necessary to obtain latent brain function by extracting out the HRF from the BOLD signal. In task related fMRI, a neural response can be extracted with explicit task inputs [26]. Unlike task related fMRI, resting-state fMRI needs to extract a neural reaction without explicit task inputs. We applied the blind-deconvolution method to apply resting-state fMRI deconvolution first published at [27].

The problem with resting-state fMRI data deconvolution is that both voxel-specific HRF and underlying neural responses are unknown. To apply blind deconvolution in resting-state fMRI data, we assume that the resting-state signal is related to spontaneous events. The goal is to extract the HRF from those pseudo-events. The pseudo-events are detected by passing a given threshold on the observed BOLD signal. Then the estimated HRF is constructed by fitting it to a double gamma function and two-time derivatives. The final neural response signal is reconstructed by Wiener deconvolution with the estimated HRF [27].

1.4 RSFC in psychiatric research

As mentioned above, resting-state fMRI can reveal some fundamental characteristics of brain function. Using the resting-state spontaneous fluctuations measured by fMRI, the connectivity analysis aims to model the operational interactions between spatially distinct brain regions [28] [29]. There are two common types of connectivity: functional connectivity (FC) and effective connectivity (EC). FC studies the bidirectional and simultaneous activity between two regions of interest (ROIs) while EC studies one directional causal information from one source ROI to a target ROI. One straightforward technique to construct an FC matrix is the seed voxel correlation mapping, which can be achieved by calculating the Pearson's correlation coefficients

between each fMRI signal across different ROIs, resulting in a parametric image [30]. In the research area, there is a significant correlation between RSFC and functionally integrated neural networks in the human brain comprised from posterior cingulate, ventral anterior cingulate, and ventromedial prefrontal cortex [31]. Above this finding, studies have also explored many resting-states functionally connected circuits related to motor, language, and visual networks [31] [32].

The combined technique of rs-fMRI data and functional connectivity mapping in psychiatric disorder research provides a series of interesting findings [33]. Autism spectrum disorder (ASD) is a highly heterogeneous neurodevelopmental disorder with various behaviors. Those behaviors are associated with abnormal patterns of RSFC, including the reduced correlation between frontal and posterior brain networks [34], and local FC strengthens with long distance FC reduction [35]. A novel study found that incorporating vertices on the cortical surface into the DMN and sensorimotor networks are highly related to the severity of ASD symptoms [36] [37]. Regarding schizophrenia research, an investigation found that the RSFC in the cortical network and cortical-subcortical is changed in schizophrenia [38]. This result shows that cortical association networks in patients with schizophrenia have reduced RSFC within the fronto-parietal network and lower segregation between the fronto-parietal network and DMN. Similarly, compared to healthy control subjects, AD patients reveal reduced RSFC between the right posterior insula and left anterior cingulate, and reduced RSFC between the right amygdala and right secondary somatosensory cortex [39]. In summary, RSFC becomes an applicable tool in classifying and diagnosing various types of psychiatric disorders. The only issue is that the statistical results are inconsistency across different studies [33]. Our initial idea is to apply a deep learning approach in analyzing RSFC data to reduce the effect of experiment and individual differences.

1.5 Organization of the Dissertation

The first chapter presents a concise introduction to fMRI, preprocessing, and resting-state functional connectivity. Chapter 2 reviews some foundational techniques related to machine learning and deep learning, including supervised and unsupervised learning. Chapter 3 presents the transfer learning approach to classifying autism using the Autism Brain Imaging Data Exchange (ABIDE) resting-state fMRI dataset. We used Human Connectome Project (HCP) dataset as the source data in the transfer learning project. Chapter 4 presents the semi-supervised learning model we applied and the results we achieved so far in domain adaptation and psychiatric disorder prediction tasks. Chapter 5 concludes the main findings, discusses some limitations and proposes future work.

Chapter 2

General Methods

2.1 Review of deep learning methods

2.1.1 Machine learning and Deep learning

In the previous decade, machine learning, as one of the most important artificial intelligence techniques, has been rapidly developed in many neural science studies. Compared to the conventional mass-univariate analytical techniques, ML methods are multivariate method. It takes the inter-correlation between voxels into analysis, breaking the univariate limitation of the traditional method. In addition, ML methods have been proved to achieve single subject brain disorders prediction and neural image based biomarkers identification in an extensive range of neural science studies [40]. However, some studies review that machine learning also has its limitations. Most common ML methods like support vector machines (SVM) and kernel methods do not directly perform well on raw data or end-to-end learning procedures. It requires a level of subjectivity to manually reduce the redundancy in raw data by objectively extracting crucial information as input features to the ML model [41] [42].

Additionally, the reliable performance of machine learning on neuroscience studies applied to a limited number of participants [40]. As the input dataset is collected from different source sites, the classification performance of ML methods also decreased significantly [43] [44]. Thus, the generalizability and objectivity of ML methods remains challenge in brain disorder clinical applications.

Inspired by the human brain neural network, deep learning as a subfield of machine learning, has been developed as a less experimental dependent and more data-driven approach.

Despite the varying degree of success of deep learning techniques in fields of computer vision, speech recognition, and natural language processing [45] [46] [47], several review papers reported that various deep learning techniques have been outperformed machine learning methods in psychiatric research [48] [49] [50]. The reason includes the data-driven deep multi-layers architecture in deep neural networks is more suitable for learning order and nonlinear relations among input features and high complexity of discriminative patterns in neuroimaging data [42] [51] than traditional ML methods. The deep learning technique enables learning more generative features than the typical machine learning method because it depends on an extensive training sample size. In addition, the complex architecture of deep learning neural networks increases the level of invariance to shift. The research of convolutional neural networks (CNN) has shown that the deeper convolutional layer in the neural network can extract abstract discriminative features of patients with Alzheimer's disease [52]. Another study found that the latent representations are highly task-specific from deeper hidden layers in sensorimotor tasks [53]. Thus, in recent years, deep learning methods have become a new frontier in neuroimaging research.

2.1.2 Overfitting

Through the development of the deep neural network in early psychiatric diagnosis, overfitting has arguably become the first issue that constrains the performance of the deep neural network in this field [48]. Overfitting is a well-trained deep neural network that can achieve high prediction accuracy in a training dataset but results in poor prediction accuracy in an unseen test dataset. One primary intrinsic causation of overfitting is coupling with the small sample size and relatively high dimensionality in a neuroimaging training dataset. Increasing the complexity and capacity of deep neural networks to increase its power can also result in overfitting. Thus, some

commonly used advanced deep neural networks such as VGGNet [54] and Google Inception Network [55] were rarely applied on real fMRI data. These very deep convolutional neural networks are susceptible to overfitting when the size of the training samples is relatively small. In fMRI research, researchers try to aggregate training datasets from multiple different sites [56] [57] to increase the sample size (and guard against overfitting), given that scanning is expensive and acquiring large amounts of data on a single scanner is prohibitive. Several machine learning or deep learning studies can approach high accuracy in single-site data classification, resulting in poor accuracy in independent site classification tasks [58] [59]. Different sites of datasets result in varying experiment protocols, subject distributions, or even other data processing methods, which are crucial in prediction bias causation. Thus, prior researches indicate that the multi-site dataset requires a higher generalization of the DL method to extract independent and transferable patterns. Different ways have been developed to solve this issue and train deep neural networks by public datasets.

The most typical way to prevent overfitting is to either increase the sample size (or make the sample more representative of the general population) or use regularization. In the previous research, most of the studies built a sparsity model against the overfitting in high dimensional neuroimaging datasets, thereby providing an efficient way to enhance the generalizability in deep neural networks. However, the prediction accuracy of classifiers is poor because of the reduction in neural networks' power. Thus, there is an intrinsic trade-off between the accuracy and generalizability of prediction models. The most commonly applied methods can be used to sparse the CNN model are L1-regularization, L2-regularization, and drop-off layers [60]. Many studies have applied these methods to analyze fMRI datasets and controlled the sparsity of model in various ways [61] [62] [63] [64]. These studies demonstrate that building a sparse model can

benefit in relieving overfitting. Another commonly used method of regularization is dropout layers. The idea of a dropout layer is relatively straightforward; it gives each neuron in DNN a probability of being temporarily deactivated in training. As a result, only a partial and minor version of DNN is applied during each training iteration, so the complexity of DNN is significantly reduced [65].

2.1.3 Prediction Interpretation

Biomarker identification in the deep learning model are another difficult but critical challenge in deep learning study. Unlike the statistical group analysis method, in most previous deep learning research, how the deep learning model classifies and predicts the results remain unknown. In psychiatric disorder diagnosis clinical approaches, the interpretability of the deep model is crucial. The mechanism of the deep learning model needs to be further clarified and verified. In a transfer learning study [66], researchers applied PCA to analyze the first layer and contractive auto encoder (CAE) to analyze the third layer, followed by PCA for the fourth layer. The methods are complex and lack of consistency at this point. Other studies compared the weights in a specific layer and end up with a feature variance map such as a brain network figure [66] [67] [62]. This method may not be sufficient to validate how a well-trained deep neural network predicts unseen data that propagates forward through all layers and further determines which biomarker contributes the most in final decision making. In this study, we used a recently developed method to either identify discriminative biomarkers in mental disorder classification or assess the robustness of RSFC features in different datasets.

2.2 Large-scale online datasets

ABIDE I&II

Autism spectrum disorder (ASD) remains a challenge in over 1% of children, but diagnosis and prediction are sufficient for earlier ages. The Autism Brain Imaging Data Exchange (ABIDE) initiative has large-scale functional and structural brain imaging data to analyze various types of ASD and respond to the clinical requests (http://fcon_1000.projects.nitrc.org/indi/abide/). There are two substantial collections of datasets, namely, ABIDE I [68] and ABIDE II [69]. Specifically, ABIDE I (http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html) was released in August 2012, which aggregated fMRI data from 18 international sites, including 539 ASD patients and 573 healthy controls. ABIDE II (http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html) was initially developed to focus on the study connectome of the ASD brain network. Later than ABIDE I, it was released by June 2016. ABIDE II involves 19 acquirement sites, including 1114 subjects from 521 individuals with ASD and 593 healthy controls, with subjects' ages ranging from 5 years to 64 years.

Human Connectome Project

The Human Connectome Project (HCP; http://www.humanconnectomeproject.org/) aims to construct an unparalleled compilation of the complete structural and functional neural connections within and across subjects. It shares large-scale data, including resting-state fMRI data, to address human connectional questions. In practice, we collected the rs-fMRI data from the HCP S1200 release [70]. It contains 1206 healthy young subjects aged 22 to 35 years, resulting in 1094 anatomical and functional rs-fMRI data. The anatomical and functional data were acquired by Siemens 3T scanner at Washington University at St. Louis for each subject.

Anatomical MRI data are 0.7mm isotropic, and the functional MRI data are 2mm isotropic. The TR is equal to 0.72s. It has one thousand two hundred frames per run (14.4 minutes) and four runs in two sessions.

1000 Functional Connectome Project (1000 FCP)

As the essential collection of ABIDE data, the 1000 Functional Connectomes Projects is a collection of rs-fMRI datasets from over 1000 subjects acquired in more than 30 independent global studies. The large and heterogeneous sample of rs-fMRI data can facilitate the analysis of consistencies in the default brain network across different subjects and institutions [71]. Given that all the sites and studies are independent, the between-study heterogeneity remains a challenge in the homogenous analysis of the 1000 FCP dataset. In 2010, one original study [72] developed a method to use heterogeneity data as a whole result in expected homogenous results.

ADNI

The Alzheimer's Disease Neuroimaging Initiative (ADNI) (www.loni.ucla.edu/ADNI) [73] gathers various types of data that aim to improve clinical and research studies to predict and treat Alzheimer's disease (AD). It includes 63 sites from the US and Canada to track the progression of AD in neuroimaging and biochemical data. ADNI helps researchers develop a standardized protocol to analyze multi-site shared AD data. In the recent decade, ADNI has become an entire database in diagnosing AD and predicting future AD. Many studies have achieved over 95% accuracy and biomarkers in AD early prediction from both structural MRI data or functional MRI by using proposed feature extraction and classification methods [74] [75] [76] [77].

AOMIC

The Amsterdam Open MRI Collection (AOMIC) provides multimodal (3T) MRI data, including T1-weighted, diffusion-weighted, and functional MRI data [78]. All healthy participants have details of demographic information. This study collected two different releases named PIOP1 and PIOP2 (Population Imaging of Psychology) from the AOMIC resting-state dataset.

HBN

The Child Mind Institute launches the Healthy Brain Network (HBN), and its focus on sharing a collection of samples from over 10k New York area child participants. The HBN data includes diffusion MRI, morphometric MRI, resting state fMRI, etc. We used resting-state fMRI and structural MRI data in this study. The demographic information, protocol and experiment design of HBN database can be found in [79] .

There are many other large public fMRI datasets that are currently available such as the UK Biobank [80] (N>40,000) and the adolescent brain cognitive development (ABCD) [81] (N=1,1975) etc. Those databases have fMRI data acquired from subjects with varied ages, gender, race, and other demographics that can enhance the generalizability of the training model.

2.3 Methods

2.3.1 Supervised learning models

Multi-layer Perceptron (MLP)

Multilayer perceptron (MLP) is a vanilla class of feedforward neural network. It consists of three major components: an input layer, a hidden layer, and an output layer. The hidden layer in MLP with a learned nonlinear activation function can transform input data into linearly separable space, making MLP a logistic classifier. MLP commonly uses a supervised training method called backpropagation, which optimizes and updates the parameters from output layer to input layer. MLP is also referred to as fully connected layers (FNNs), where each node is connected with all other nodes in each layer. Given that excessive hidden neurons are involved, overfitting remains a critical issue in FNN like architecture [82]. In RSFC studies, MLP has often been applied as a classifier to predict behavior and demographics [80] and used in comparison to more advanced neural network architectures.

Convolutional Neural Network (CNN)

Unlike MLP, convolutional neural network (CNN) aims to explore the input of a 2D/3D image/volume or matrix instead of a single vector (Fig.1). One property of most images is that the adjacent pixels share similar information, making CNN capable of reducing the number of parameters compared with FNN architecture. The essential component of CNN is a convolutional layer. One convolutional layer consists of a series of learnable filters in in-depth dimension. All filters can be seen as a partial kernel to filter raw images simultaneously. In the forward learning process, each filter takes a 2D convolution calculation across the width and height of the input image and pass the result to an activation function. After nonlinear activation,

a weighted stack of weighted activation maps is created. In this way, a specific pattern of features can be detected and located by a specific filter. In other words, every entry in the result can be interpreted as an output of neurons with one particular image location. They share parameters with neurons in the same activation result [83] [84]. This local weight sharing significantly reduces the number of learnable parameters and reduces the model's capacity. So, it is a suitable property against overfitting. CNN also consists of a pooling layer and fully connected layers. All layers are trained by feedforward backpropagation.

Early studies exploited the idea of using CNN representations in performing prediction and showed good within-subject performance [85]. As a result of the character of a 2D feature extractor, many studies applied CNN as a decoder to analyze the fMRI signal from the visual cortex [86] [87] [85] [88]. Previous studie validated the reliability of the established mapping between fMRI data and convolutional layers to extract robust feature representation of visual stimuli. So, it can be used as a generic mapping method in visual signal decoding and encoding [86]. CNN is also widely applied in identifying discrimination of schizophrenia and ASD [89] [61].

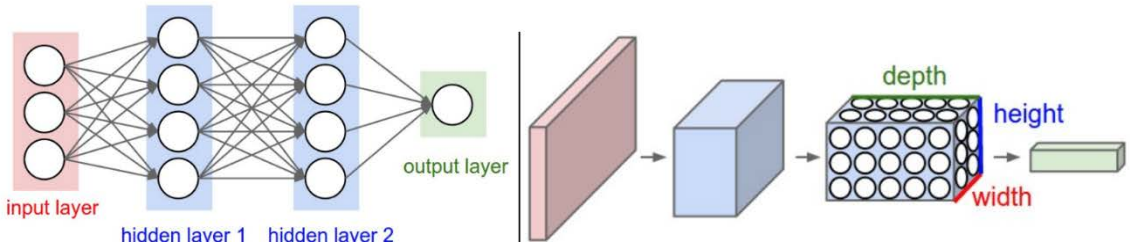

Figure 1. An architecture overview of CNN. Left: A regular 3-layer neural Network. Right: A ConvNet arranges its neurons in three dimensions (width, height, depth), as visualized in one of the layers. Every layer of a ConvNet transforms the 3D input volume to a 3D output volume of neuron activations. In this example, the red input layer holds the image, so its width and height would be the dimensions of the image.

2.3.2 Unsupervised learning models

Different from supervised learning method, unsupervised learning aims to learn undetected patterns in a dataset with no observed labels and minimal human supervision. In contrast to learn human labeled data in a supervised learning method, unsupervised learning focuses on modeling probability densities over inputs. One of the most typical use of unsupervised learning is clustering.

Autoencoder

Autoencoder (AE) is one of the typical unsupervised learning methods. It learns efficient data coding from unlabeled data in downstream tasks [90]. As illustrated in Fig 2, it consists of two main processes, encoding, and decoding. In the encoding process, also named the reduction phase, an encoder reduces the dimension of raw data and extracts the efficient feature representation. In the decoding process, the reconstruction phase, a decoder reconstructs the estimated input from latent features in the encoding process (Fig.2). Both encoder and decoder are trained simultaneously to achieve optimal results. By penalizing the network with the

reconstruction error, our model can learn the essential attributes of the input data and how to best reconstruct the original input from an "encoded" state. Ideally, this encoding will learn and describe latent characteristics of the input data.

AE has been widely used in offline pre-training models to extract feature representation from high dimensional fMRI datasets because of its dimension reduction ability [91] [92] [93]. The basic idea to pre-train a model is to combine an autoencoder (AE) or deep belief network (DBN) as an unsupervised learning model to train an unlabeled dataset and an adaptive model for specific classification or regression tasks in supervised learning. The result indicated that a pre-trained AE could effectively reduce the heterogeneity among the input datasets from different sites [94]. In a previous study [95], the author found that the error rate of schizophrenia classification significantly decreased from 20.2 (± 1.2) to 14.2 (± 0.4) if the weights of the deep neural network (DNN) are initialized by a well pre-trained stacked autoencoder (SAE) than without it. The researchers controlled the sparsity of SAE by the L1-norm regularization term. One ASD diagnosis paper employed a similar denoising SAE to pre-train their DNN, resulting in over 70% accuracy in identifying ASD versus control patients in the dataset [96]. They also identified the positive and negative functional connectivity (FC) features in predicting ASD from the resting-state fMRI dataset. One study combined DNN with a pre-trained SAE as a feature selection model, resulting in a 9.09% improvement in the ASD classification task [97]. Other work also proved that prior learning experience from a pre-trained model could increase the generalizability in the feature extraction process and complex pattern identification in mental disorder studies [59] [91].

One advanced type of AE called denoising autoencoders (DAE) was proposed to train the predictive model for better generalization. The estimated classification could improve novel test

data accuracy and alleviate the overfitting effect. The decoder of DAE reconstructs input data based on a corrupted version of input [94] . In RSFC research, some random features of the input RSFC data are initially set to zero at first. In this way, the uncertainty and diversity of input data have been increased, allowing pre-trained AE to have greater accuracy in predicting novel data [98]. Especially in a multi-site dataset, one study tested their DAE model by leave-one-site-out classification. They found that 13 out of 18 sites of the ABIDE multi-site RSFC data have higher prediction accuracy than global results obtained by combining data from all sites [96].

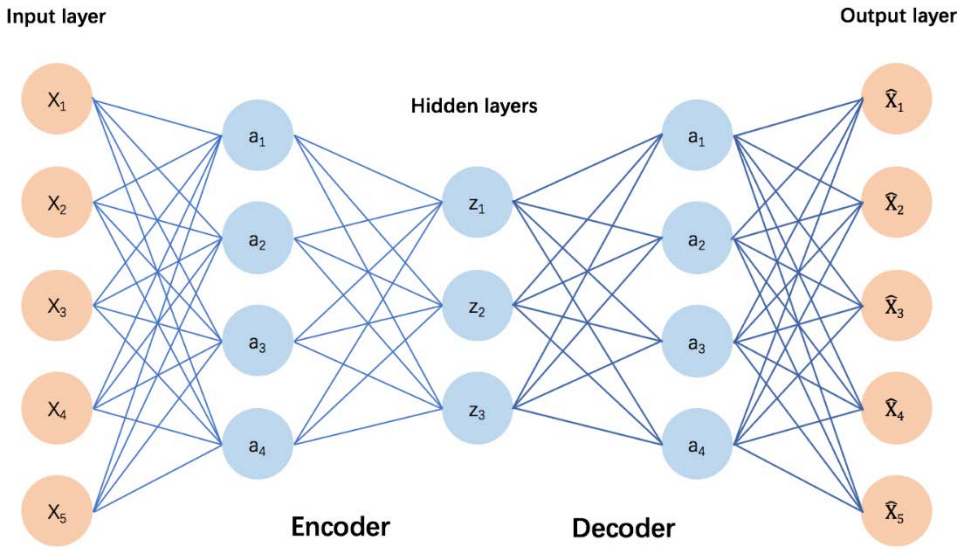

Figure 2. A simple autoencoder consists of an encoder and decoder. Encoder proposes to extract latent representation *z* of data *x* from the input layer to hidden layers, and decoder proposes to reconstruct data from latent feature *z*. The reconstruct error is the difference between the original input data *x* and the consequent reconstruction data $\bar{x}$.

2.3.3 Model Explanation methods

Given the complex and non-linear mechanism of the deep learning model, it is often treated as a black box to predict the result. However, in the mental disorder classification task, interpreting the deep learning model's decision-making becomes an important issue for further

clinical application. Several methods were proposed in interpreting the classification result making process including IntegratedGradients [99], Layer-wise Relevance Propagation (LRP) [100], DeepTaylor [101], and DeepLIFT [102], etc. In theory, there is no conclusion that which one interprets model best. Among them, IntegratedGradients computes the partial derivative of the output to each input feature. Different from the common gradient-based method, the gradient derivative is the mean gradient while the input change from a baseline $\bar{x}$ to x. The baseline $\bar{x}$ is normally set to zero. As illustrated in Fig.3, LRP constructs quantity of relevance of unit x in layer l referred as r_i^l, through a backpropagation from output units to the input features. There are multiple propagation rules to calculate $r_i^{(l)}$ in LRP, such as Basic Rule (LRP-0), Epsilon Rule (LRP- ϵ) and Gamma Rule (LRP-γ) [103]. Similar to LRP, Deep LIFT also proceeds in back pass computation. Different is, the attribution of each unit is assigned as a comparison between the relative effect of the original network to the new network with reference baseline input. Normally the baseline is set to zero. All methods discussed to construct a result heatmap at the end, which has an attribution score of all features relative to a target prediction unit. In psychiatric classification tasks, interpretation methods were proposed to construct a relevant map of all ROIs in the human brain and further identify the biomarker in the disease group.

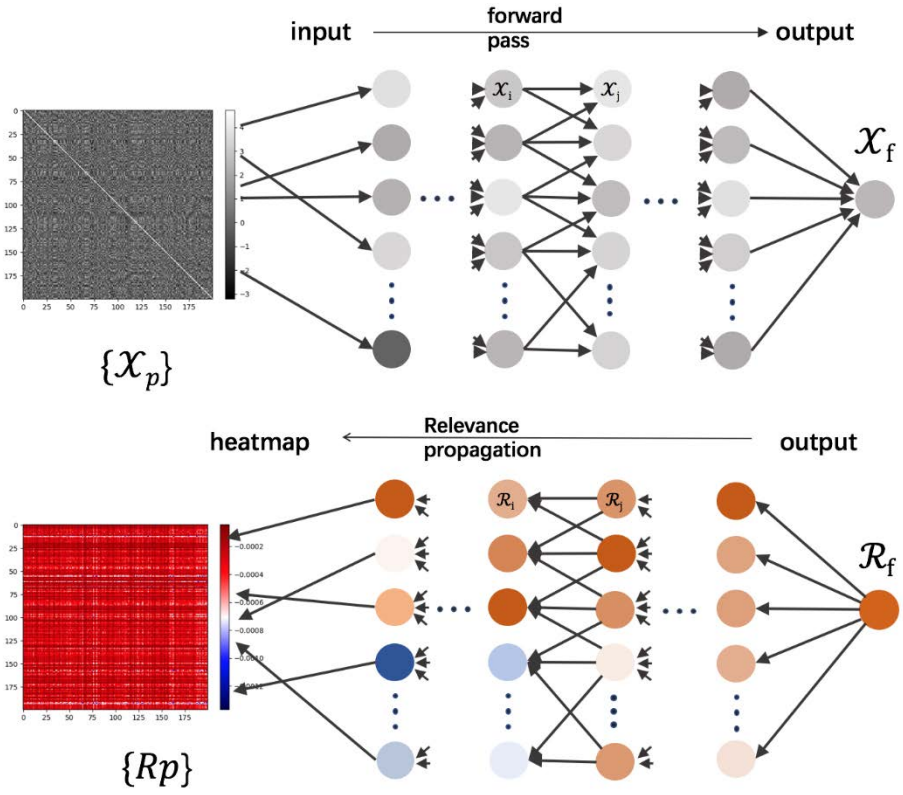

Figure 3. Computational flow of LRP. A prediction for the class ASD is obtained by forward-propagation of the pixel values $\{x_p\}$, and is encoded by the output neuron x_f. The output neuron is assigned a relevance score $R_f = x_f$ representing the total evidence for the class ASD. Relevance is then back-propagated from the top layer down to the input, where $\{R_p\}$ denotes the pixel-wise relevance scores, that can be visualized as a heatmap.

2.3.4 Domain Adaptation

In order to understand domain adaptation, consider a scenario wherein an algorithm needs to perform a classification task on the Dutch language. One could improve its performance by first training it on a larger database of English and then training on a relatively smaller database of Dutch. Even though the two languages are different, at least some of the rules of language are similar. Therefore, the lack of availability of a large Dutch database can be alleviated by learning some linguistic rules from a larger English database. This improves algorithmic performance when

some linguistic rules learned from English may be equally applicable in Dutch. This is the essence of domain adaptation [104]. As illustrated in **Figure 4**, the data distributions are different in the source domain and target domain although the two groups are separable in both the domains taken independently. However, the classifier learned from the source domain (the red dotted line in **Figure 4a**) cannot directly be transferred to the target domain (**Figure 4b**). This affects the generalizability of the classifier. Thus, the objective of domain adaptation is to learn the differences in data distributions and improve the target domain classifier (black dotted line in **Figure 4c**) by jointly optimizing the classification and domain fusion (illustrated by approaching and splitting arrows in **Figure 4c**) [105]. In neuroimaging research, the transductive scenario assumes that the dataset from the source domain has annotated labels from an expert and the dataset from the target domain may not have labels. The domain adaptation approach is jointly optimized to minimize the domain shift effect across source domain data and target domain data [106].

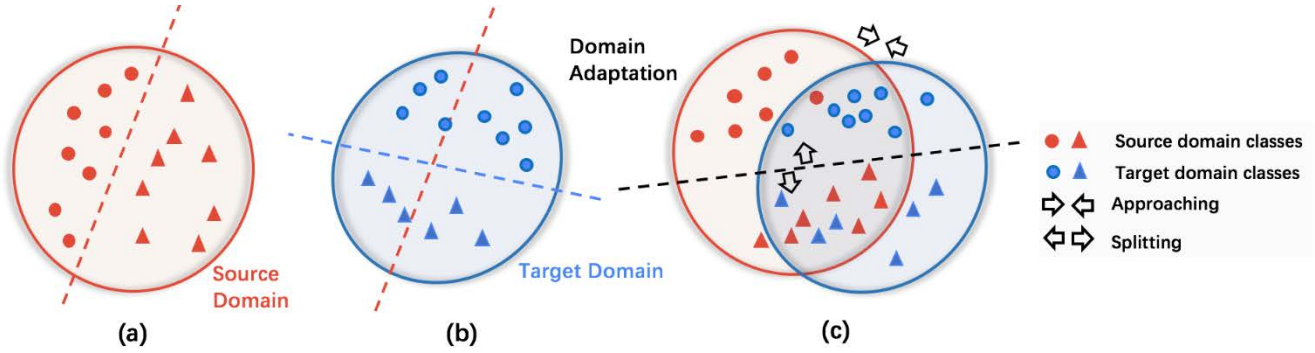

Figure 4: *(a) and (b) are the classifiers before domain adaptation in the source and target domains, respectively. The domain adaptation process aims to reduce the domain divergence by maximizing the domain confusion as well as minimizing the classification loss in (c). The red line in (a) illustrates the decision boundary from training a source domain classifier. When this is transferred to the target domain as is, it is sub-optimal. The blue line in (b) illustrates the desired decision boundary from training a target domain classifier. The black line in (c) illustrates the decision boundary result from training a domain adaptation classifier.*

Chapter 3

Transfer learning for improving neuroimaging based diagnostic classification

Abstract

Overfitting, the main issue that constrains the validity and generalizability of machine learning in neuroimaging based diagnostic classification, is in part due to small sample sizes in relation to what is required for generalization. Even with data aggregation (such as in Autism Brain Imaging Data Exchange or ABIDE), the relatively smaller sample sizes are a result of the fact that it is difficult/expensive to acquire data from clinical populations. With healthy controls, we have comparatively larger samples available. Therefore, we propose to address overfitting by using larger healthy samples (from Human Connectome Project or HCP) to learn the diversity of neural signatures of healthy controls, with the aim of transferring that learning into the context of discriminating autism from healthy controls in ABIDE. In order to do so, we developed a complete variational autoencoder based transfer learning framework including data oversampling, model pre-training, classifier training and testing, and model explanation. Then, the performance of transfer learning was estimated and visualized. The transfer learning classification model achieved about 7% more accuracy on site-mismatched data than obtained without transfer leaning. Overall, we have demonstrated the applicability of transfer learning within a deep learning framework for utilizing larger samples of available healthy control data to improve generalizability and accuracy of diagnostic classification in ASD, as well as reduce the harmful effects of inter-site variability on classification. We believe the proposed framework is potentially applicable to other disorders as well.

Keywords**:** *functional connectivity, autism spectrum disorders, transfer learning, variational autoencoder, convolutional neural network*

3.1. Introduction

The use of artificial intelligence in diagnosing psychiatric disorders based on neuroimaging data is an attractive proposition since it provides an element of objectivity in an otherwise traditionally subjective mode of diagnosis. Despite a strong logical case, its practical implementation and utility remains a challenge since the prediction accuracy is limited and generalizability is uncertain. Here we focus on one such mental disorder, autism spectrum disorder (ASD), which is a complex neurobehavioral disorder that is clinically diagnosed by impairment in communications and social skills [107] [108] [109] [40]. In recent years, deep learning (DL) has been developed as a data driven approach to build hierarchical multi-layer structures to automatically predict outputs of a given system from novel inputs [110] [111].

3.1.1. Limitations of existing deep learning approaches

Despite the varying degrees of success of DL techniques in fields of computer vision, speech recognition, and natural language processing [112] [46] [47], its applicability in the context of neuroimaging based diagnosis of mental disorders is yet unclear, primarily due to questions about overfitting and generalizability. For example, using open-source large data such as Autism Brain Imaging Data Exchange (ABIDE) [58], several papers reported that various DL techniques outperformed traditional machine learning methods in ASD classification research [113] [49]. However, overfitting is arguably the first issue that constraints the performance of deep neural networks in this field [114] [115]. Overfitting refers to a well-trained deep neural network that can achieve high prediction accuracy in the training dataset but results in a poor prediction accuracy in an unseen test dataset [116]. One major cause for overfitting is a combination of the small

sample size and relatively high dimensionality in the training dataset, which is also known as the curse of dimensionality or small-*n*-large-*p* problem [117].

Another issue limiting the performance of deep neural networks (DNN) is the heterogeneity between training and test dataset, especially when datasets are collected from different sites as in large open-source databases. Even though previous studies have used relatively large training samples, the prediction accuracy from multisite data was significantly lower than that from single site data [118] [59] [115]. For large open-source data such as ABIDE I, the data are pooled from 18 acquisition sites with different scanner types, acquisition parameters, age/gender distribution of subjects [119]; these lead to different data distributions from different sites. However, in real clinical circumstances, test data will probably be acquired from different institutions than the ones on which the classifier is trained. A classifier cannot easily predict novel data distributions based on different data distributions in the training dataset. A prior study using ABIDE data has reported that classification accuracy can be significantly degraded when training and testing data are drawn from different sites as compared to them being drawn equally from those sites [115]. Thus, prior research indicates that the multi-site data requires DL classifiers to be able to generalize better [120] [59].

3.1.2. Transfer learning in neuroimaging

To address the issues of overfitting and ensuing lack of generalizability across sites in multi-site data, we propose transfer learning [121]. One major cause of overfitting is the limited sample size and heterogeneity in the proposed datasets. A straightforward solution to reduce overfitting is combining the knowledge learned from datasets in other source domains and fine-tuning an adaptive neural network by training fewer examples from the current target domain. This principle of “transferring” knowledge learned from one dataset into another is known as transfer

learning and is a dominant approach in multiple fields of science and engineering [122] [123], and has the potential to be applied in neuroimaging based diagnostics. It may be an effective technique to enhance the robustness and generalizability of artificial neural networks with complex neuroimaging datasets. The utility of such transfer learning (TL) has been demonstrated in unrelated contexts and, to a limited extent in disorder classification from neuroimaging data. In TL, the source domain is defined as the space from which initial learning is derived from, and the target domain is defined as the space into which this initial learning is transferred to.

In neuroimaging, TL is especially relevant because we now have relatively large datasets with data from healthy controls. For instance, the initial release of UK biobank has 10065 subjects with both structural and rs-fMRI data [124]. Human Connectome Project (HCP) comprised 1206 healthy young adults (age 22–35). There were 1094 subjects with both structural MRI and rs-fMRI. [125]. Amsterdam Open MRI Collection (AOMIC) [126] includes three datasets as well as detailed demographics and psychometric variables from a relatively large set of healthy participants (N=928, N=226, and N=216). Yet, acquiring such large amounts of data from every disease population remains difficult and expensive. Therefore, the idea is that if we are able to use the large amount of healthy data to learn neural representations of mental health, then such learning to can be "transferred" to a context where individuals with a mental disorder are being distinguished from controls. In short, when characterizing neural representations of mental health disorders, knowing the neural correlates of health precisely, allows us to leverage the large amount of data from healthy controls to our benefit.

Previous neuroimaging studies have used TL in different contexts [127] [128]. One study [129] applied a novel two-path 3-D convolutional neural network (CNN) architecture, to train structural MRI and functional MRI simultaneously. They achieved an accuracy of 69.15% in an

attention deficit hyperactivity disorder (ADHD) classification task using a multi-modality 3-D CNN model with TL between modalities. TL has also been applied in the early diagnosis of Alzheimer's disease (AD) [130] [131]. In this case, although TL is not from a different dataset but rather from a subset of the same dataset, they obtained better accuracy in multi-site diagnostic prediction [132] with TL compared to without. Vakli et al. used resting state functional connectivity (rs-FC) to predict age in healthy individuals and found that accuracy and generalizability of the model on their in-house dataset improved when they used a model that had been pre-trained using larger public databases such as Nathan Kline Institute Rockland Sample (NKI-RS) dataset [133].

In contrast to these earlier studies, we are proposing to "transfer" the learning from healthy control data (obtained from HCP [134]) to situations where one has to predict whether a given subject is healthy or has a mental disorder (ASD in our case). The target domain (in our case, ABIDE data) is defined as the space to which the classifier pre-trained in the source domain (in our case, HCP data) is applied for classification. For the target domain, we utilized data from subjects with ASD and healthy controls available in the ABIDE dataset [135], pre-processed with a pipeline identical to that used for HCP data.

3.1.3. Data oversampling using generative models

Even though TL can address overfitting and enhance the generalizability of model, the class imbalance issue caused by TL cannot be ignored. Class imbalance arises when a dataset has large samples in one major class, and relatively smaller samples in other classes. Different from computer vision, natural language processing or other fields, publicly available large neuroimaging databases containing healthy control participants are available as compared to databases containing subjects with mental disorders. For example, ABIDE I and II [58] (N~2000) are available for ASD

and the Alzheimer's Disease Neuroimaging Initiative (ADNI) [130] (N~3000) for Alzheimer's. In contrast, several datasets are available for healthy controls including HCP [134] (N~1200), Healthy Brain Network (HBN) [136](N~ 2500) and UK Biobank [124] (N>10000). Pooling these will provide us with over 10,000 healthy control subjects, which is an order of magnitude more than the number of subjects available for any given mental disorder. Therefore, any transfer learning from healthy datasets (as explained in the previous section) would create a scenario of unbalanced datasets during classification [137], wherein the DL model is trained by many more healthy control class samples as compared to patient class samples. This biases the model towards features from the major class, thereby impacting the overall error rate [138] [139] [140]. This situation is aggravated in TL and needs to be addressed.

To balance the dataset, the commonly used resampling techniques include undersampling the majority class data [141] or oversampling the minority class data [142]. Generally, the undersampling works better than oversampling [143], but the latter does not lead to loss of data from the original dataset. The most widely used oversampling approach is synthesizing new examples and is called Synthetic Minority Oversampling Technique (SMOTE) developed by Chawla et al. [142]. The major steps of SMOTE include selecting samples closed in the latent feature space, drawing a separate line between the samples and then synthesizing new samples at a point along the line. However, SMOTE is known to be less effective on high dimensional data such as gene expression microarray data [144]. Therefore, it is unlikely to work well on high dimensional fMRI data. Different from the traditional SMOTE method, with the development of generative modeling, Variational Autoencoder (VAE) [145] and Generative Adversarial Networks (GAN) [146] for balancing classes have been successfully applied across various fields [147] [148] [149] [150]. The details of VAE will be introduced below. In short, a generative model can learn

the true underlying distribution of input observations in latent space during training, and can subsequently be applied to generate synthetic observations that closely resemble those of the training dataset [151]. By comparing different approaches, Fajardo et al. [151] found that a VAE based model outperformed GAN on benchmark MNIST (Modified National Institute of Standards and Technology database) image dataset. Wang et al. [152] developed a novel generative VAE model to generate synthetic data as a data augmentation strategy to achieve state-of-the-art 90.4% accuracy on high resolution ImageCLEF-DA images, and also reduce the divergence of input data distributions. Therefore, in this study, we utilized a VAE model to generate synthetic FC features from minority ASD class, and then use the synthetic data to balance the large-scale heathy control samples transferred from the HCP database.

3.1.4. CNN-based VAE model

Unlike conventional autoencoders, VAE is a generative model that can train a generative representation from input data and generate new samples such as text, image, or language from the latent sample distribution [122] [123] [145]. Learning a generative representation from input data could potentially aid in improving generalizability of a CNN, something that we aim to achieve using TL. However, applications that use VAE as an unsupervised learning model in TL remains limited. One group used VAE as an encoding model to analyze visual cortex fMRI data [153]. They found that the VAE model decodes video reconstruction activity in a more convenient way by transforming fMRI activity into low dimensional VAE latent space representations. Besides that, the application of VAE in neuroimaging is in its infancy, mainly because most fMRI studies have focused on learning a compressed representation of input data while VAE learns the parameters of a probability distribution representing the data [145] [127]. Since VAE learns to model the data, researchers can sample from the distribution and generate new input data samples.

This provides distinct advantages compared to the traditional approach generally used in neuroimaging based classifiers. Different from abovementioned studies, the goal of pre-training the model in the proposed TL approach is not just learning a compressed version of fMRI data, but also to learn the probability density across different data distributions because we have fMRI data from different domains (or sites). Thus, the application of VAE in TL as applied to fMRI data is among our novel contributions.

For ASD classification, we applied a CNN model driven by FC features as discussed in Zou et al. [129]. The data-driven CNN has a lower dimension and less parameters than normal CNN, and hence is considered robust against overfitting in FC classification [154] [155]. CNN has also been used in Alzheimer's classification, and has been proven to be an advanced supervised learning model useful in classifying Alzheimer's from controls in the ADNI data [130]. To implement a complete TL model, we transferred the encoder from VAE to CNN and replaced the decoder with fully connected layers for classification. Similar to prior fMRI research, the FC adjacency matrix was treated as a normal square image input to the designed convolutional filters.

Although previous studies have used TL to develop state-of-the-art learning models in different contexts [127] [128], most of these studies have not validated the effect of the TL model they applied on classification. To validate the process of how TL can improve the prediction, we tested the relationship between the sample size of HCP data in the pre-training model and classification accuracy, with the hypothesis that inclusion of increased number of healthy subjects from HCP will improve classification accuracy in ABIDE data.

3.1.5. Stacked Autoencoder (SAE) in model pre-training

In many previous DL-based studies, various kinds of hierarchical offline pre-training models were widely used [156] [92] [157]. Considering the complicated properties of fMRI

neuroimaging data, a well-trained unsupervised learning model can extract essential features from the higher dimensional structure of neuroimaging data [156]. This is a method of feature selection for DL models, similar to feature selection in traditional machine learning models. Selecting only essential features improves the performance of the classifier. A typical way to pre-train a model is by augmenting the main classifier with an autoencoder (AE) or deep belief network (DBN) [49]. An AE consists of two major components, an encoder to construct latent feature representations and a decoder to reconstruct original input data. The model is trained by minimizing the reconstruction error. Prior research showed that the features extracted by the stacked autoencoder (SAE) are robust and efficient [156]. Another research report indicated that a pre-trained denoising AE can partially address the heterogeneity among data collected from different sites in the ABIDE dataset [94] [96].

Among the abovementioned studies, most have built a sparsity model against overfitting in high dimensional neuroimaging datasets [156] [95]. The sparsity of AE is critical in enhancing the generalizability of the features extracted. However, the drawback of sparsity is that it introduces a bias and reduces the power of neural networks. Thus, sparsity is an intrinsic trade-off between bias and overfitting of prediction models. The most commonly applied methods to increase the sparsity of CNN models are L1-regularization, L2-regularization, and drop-off layers [60]. Many studies have applied sparse deep neural networks to analyze fMRI data, thereby controlling the sparsity in various ways [61] [155] [158] [159] [95]. For example, measures of ANOVA in one schizophrenia study revealed that the effects of sparsity control in pre-trained hierarchical layers to the error rates were statistically significant (Bonferroni-corrected $p<10^{-7}$; d.f.=999) (Kim, Calhoun, Shim, & Lee, 2016). Therefore, to enhance the features extraction ability

and address overfitting, we trained sparse CNN layers with L1-regularization [160] in a stacked AE framework in this study.

3.1.6.– LRP in feature identification and interpretation

The interpretation of prediction results obtained from DL is a critical challenge in most studies attempting neuroimaging based diagnostics. Unlike statistical and traditional machine learning models such as the support vector machine, the mechanism by which the DL model classifies and predicts the result remains a black box [161] [162]. To better explain the basis of the results, previous studies using rs-FC for ASD classification [163] [164] have used the LRP (Layer-Wise Relevance Propagation) algorithm to identify the most discriminative FC features in ASD classification. The major advantage of the LRP method is that it can generate a pixel-wise scaled heat map (in the FC matrix space) to construct the contribution of each FC value towards the final classification result. In the end, on the basis of the contribution map of FC features, we can infer which neural mechanisms (based on FC) may be most predictive of ASD diagnostic status.

Various approaches may be used to interpret the results obtained by artificial neural networks, i.e., identify features that are discriminable across groups and likely led to the accuracy obtained. In some neuroimaging studies, the discriminative features are identified from the weights of shallow layers, while more abstract features are in the deeper layers of DNN [165]. Therefore, we applied LRP to run through all layers from the deepest output layer to the shallowest one. This type of feature interpretation approach is absent in TL neuroimaging studies so far. However, the LRP technique has been used in neuroimaging in other contexts. For example, a study used bidirectional LSTM to model the spatial dependencies of brain activity within and across brain slices [166]. The LRP algorithm was utilized based on the LSTM decoder to decompose the states decoding decision, and it accurately identified the physiologically appropriate associations

between cognitive states and brain activity. LRP has also been applied to explain the subject-level classification outcome in structural MRI based AD classification [165]. Multiple individual heat maps are constructed to determine their neurobiological relevance. In this study, based on the well-developed LRP algorithm in the computer vision area, we designed a compact permutation statistical analysis method to identify biomarkers relevant for ASD classification results.

Overall, we present a TL framework in this paper that addresses unmet needs in the field of neuroimaging based diagnostic classification as described above. In the following section, we provide a detailed description of the methods employed.

3.2. Materials and Methods

3.2.1. Overview

We illustrate the proposed VAE-CNN framework in Fig.5. It consists of three major components: a pre-trained model to learn the encoder from unlabeled training data, a supervised learning model to predict labels from the ASD dataset, and an interpretation method to identify features in the data that have high discriminability between ASD and control subjects. As shown in Fig.5, during model training in the source domain, we used the VAE on ASD and controls not only using ABIDE training data, but also controls from HCP data. We applied a CNN model in the target domain (ABIDE) for classification between ASD and controls, with transferred parameters initializing the pre-trained VAE model from the source domain. The classification performance was evaluated in terms of change in classification accuracy with an increasing number of healthy subjects from HCP data used during pre-training. The inputs were FC matrices from each subject. LRP was used to provide mechanistic and semantic insights underlying the classification decision.

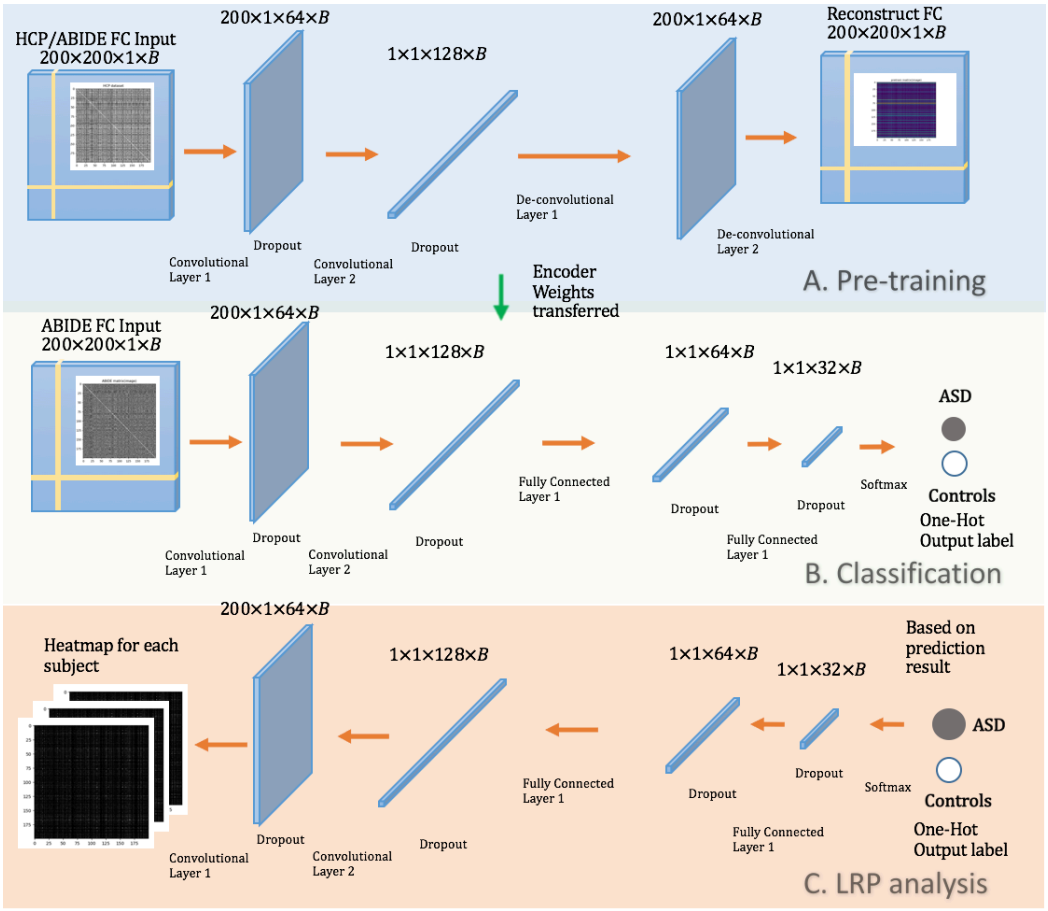

Figure 5. *The architecture of the VAE-CNN transfer learning framework, with different types of layers such as dropout layer, 1-D convolutional layer and de-convolve layer. There are three major components of this framework. (A) An example of a batch of HCP FC matrices input into the VAE model (B refers to the batch size of subjects). The offline VAE model consists of two stacked convolutional layers as an encoder, and two de-convolve layers as the decoder. (B) Then the weights of the pre-trained encoder are transferred into the FC-adaptive CNN for ASD classification. The deconvolved and reconstructed layers in the VAE model are replaced by the fully-connected classifier in the CNN model applied to ABIDE for diagnostic prediction. (C) After classifier training, we applied LRP analysis backward from identified ASD output to the input layer. The FC heat maps of each subject are constructed for FC biomarker identification in ASD prediction.*

3.2.2 Data acquisition and preprocessing

We used rs-fMRI data from the publicly available ABIDE I dataset [135] (18 sites). We employed binary classification between controls and ASD, i.e., we considered Asperger's subjects

in the dataset to belong to ASD although ABIDE has provided a separate diagnosis for Asperger's. We did so because the number of subjects with Asperger's was quite small (N=93) compared to ASD (N=339) and controls (N=556) and we did not want to use a three-way classification problem with unbalanced datasets since TL was the focus of the study. The demographic information of participants is summarized in Table.1, including the distribution of subjects across individual sites and genders. All data were pre-processed using a standard pre-processing pipeline in the DPARSF toolbox [167]. This included the removal of the first five volumes, slice timing correction, motion correction and co-registration to the standard MNI space. Nuisance signals such as low frequency drifts, motion parameters, and white matter and cerebrospinal fluid signals were regressed out. Time series were temporally filtered by a 0.01-0.1 Hz band-pass filter.

ABIDE dataset was our target domain since we were primarily interested in predicting whether a given subject has ASD or not. We split the 895 ABIDE subjects (selected after discarding a few subjects that did not meet inclusion criteria) into training and testing datasets in the following ways: (i) site matched split: subjects were drawn randomly so that training (N=705) and testing datasets (N=190) were matched on non-imaging measures, and (ii) site mismatched split: training data from 11 sites and testing data from 7 remaining sites. This split was carried out to test the robustness of the model to different sources of training and testing datasets, and hence potentially different distributions in the training and testing data. Cross-validation tends to overestimate the performance of a classifier [168], and hence, we preferred independent test data to evaluate the model.

For the source domain, we utilized the pre-processed HCP rs-fMRI data obtained directly from their website (N=1097, includes up to the S1200 release) [134] [169] [170]. Notably, ABIDE and HCP datasets are distinct on various aspects, including subject distribution, age difference,

gender difference, ASD syndrome, and pre-processing pipeline. This distinction is also a source of diversity (at least in the healthy control data) that the classifier would be exposed to via TL, potentially improving its ability to generalize to the larger population.

	Control			Autism		
	Age Avg (SD)	Gender	Count	Age Avg (SD)	Gender	Count
CALTECH	28.87(11.21)	M: 15 F:4	19	23.75(6.36)	M:10 F:3	13
CMU	26.85(5.74)	M: 10 F:3	13	26.36(5.84)	M:11 F:3	14
KKI	10.16(1.26)	M: 24 F:9	33	9.74(1.18)	M:10 F:1	11
LEUVEN_1	23.27(2.91)	M: 15 F:0	15	21.86(4.11)	M:14 F:0	14
LEUVEN_2	14.34(1.51)	M: 15 F:5	20	13.92(1.31)	M:12 F:3	15
MAX_MUN	26.21(9.80)	M: 29 F:4	33	11(0)	M:2 F:0	2
NYU	15.80(7.39)	M: 79 F:26	105	13.90(2.00)	M: 46 F:7	53
OLIN	16.81(3.49)	M: 31 F:5	36	0	~	0
PITT	18.88(6.64)	M: 23 F:4	27	18.93(7.20)	M:26 F:4	30
SBL	33.73(6.61)	M: 15 F:0	15	24.5(3.53)	M:2 F:0	2
SUDU	14.22(1.90)	M: 16 F:6	22	13.19(0.87)	M:2 F:1	3
TRINITY	17.08(3,77)	M: 25 F:0	25	16.04(2.94)	M:10 F:0	10
UCLA_1	13.25(2.11)	M: 28 F:4	32	13.10(2.62)	M:35 F:6	41
UCLA_2	12.25(1.11)	M: 11 F:2	13	12.72(1.87)	M:13 F:0	13
UM_1	14.07(3.18)	M: 38 F:17	55	12.79(2.46)	M:39 F:6	45
UM_2	16.6(3.92)	M: 21 F:1	22	14.92(1.40)	M:9 F:1	10
USM	21.36(7.64)	M: 43 F:0	43	21.08(7.78)	M:57 F:0	57
YALE	12.68(2.75)	M: 20 F:8	28	15.11(1.96)	M:4 F:2	6
Total	17.53(7.63)	M: 458 F: 98	556	16.78(6.91)	M: 302 F: 37	339

Table 1. *Demographic distribution for each site in the ABIDE I dataset*

Since the dimensionality of fMRI data is very high, we extracted mean time series from 200 regions of interest (ROIs) as defined by the popular Craddock-200 atlas [171]. These 200 functionally homogenous ROIs cover the entire brain. BOLD fMRI is a convolution of latent (unmeasured) neural activity and the hemodynamic response function (HRF). The HRF, which measures neurovascular coupling, varies between brain regions and individuals [172] [173]. This variability confounds FC estimates in healthy individuals [174] [175] as well as in subjects with

trauma [176], schizophrenia/bipolar [177] and importantly ASD [178] [179]. Therefore, we performed deconvolution on ROI time series to extract the latent neural signals [27], as in several recent studies [174] [175] [176] [177] [178] [179] [180] [181] [182] [183] [184]. FC between pairs of ROIs was then calculated as the Pearson's correlation coefficient between the corresponding latent neural signals. The whole brain 200×200 FC matrix is a weighted adjacency matrix (Fig.6), wherein each value indicates the level of co-activation between pairs of ROIs. The matrix is diagonally symmetrical because correlations do not have directionality. The diagonal values are all equal to 1 (autocorrelation) and are ignored.

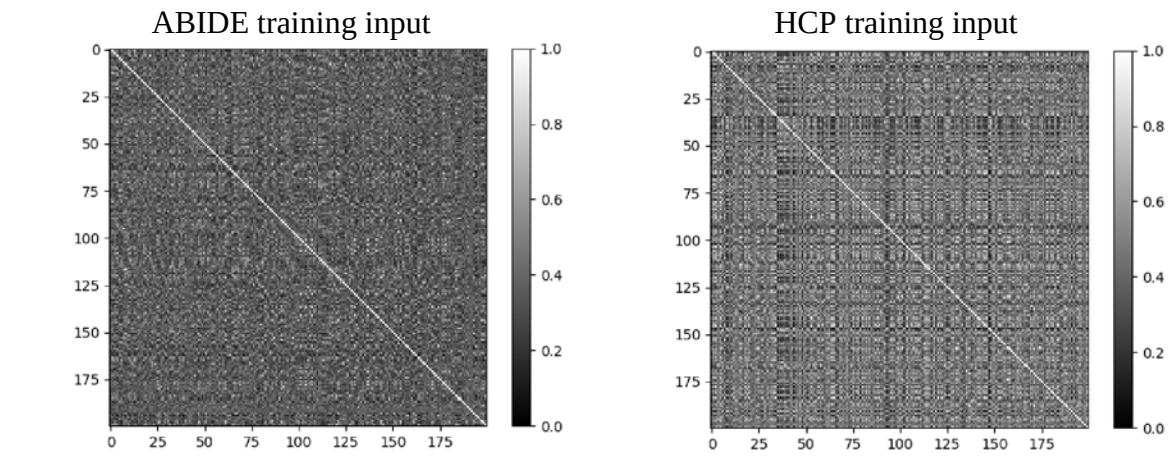

Figure 6. *Illustrative scaled images of the resting-state functional connectivity matrixes in ABIDE training dataset and HCP training dataset*

3.2.3. VAE-CNN model

VAE is an unsupervised learning autoencoder model that has recently gained popularity. Typically, VAE is applied as a latent variable generative model to learn latent feature representation from unlabeled data, and then generate new data from it [145]. Here, we pre-trained VAE as a feature extraction model without supervision. Similar to the conventional AE, VAE has

an encoder and a decoder. The input FC data are referred to as *x* and latent feature representation is referred to as *z*. The probabilistic encoder is $q_\phi(z|x)$, which is parameterized as ϕ. Given a data point *x*, it produces a Gaussian distribution over the possible values of the latent variable *z* from which the data point *x* could have been generated. In a similar way, we will refer to $p_\theta(x|z)$ as a probabilistic decoder; given a latent variable z, it produces a distribution over the possible corresponding values of x. The marginal likelihood of a data point and the loss of evidence lower bound (ELBO) $\mathcal{L}^{\text{ELBO}}$ are constructed as follows:

$$log\, p_\theta(x) = D_{KL}\left(q_\phi(z|x) \parallel p_\theta(z|x)\right) + \mathcal{L}^{ELBO}(\phi, \theta; x)$$

We need to maximize the ELBO loss, which is similar to minimizing the minus ELBO loss. Thus, the loss function of the VAE deep network can be written as

$$J^{(n)} = -\mathbb{E}_{z\sim q_\phi(z|x)}[log p_\theta(x|z)] + D_{KL}\left(q_\phi(z|x) \parallel p(z)\right)$$

where the first item is a reconstructed loss. Similar to traditional deterministic AEs, the expectation of the log-likelihood is that the input image can be generated based on the sampled values of *z* from the inferred distribution $q_\phi(z|x)$. When $q_\phi(z|x)$ is a multivariate normal distribution with unknown expectations μ and variances σ^2, the objective function is differentiable with respect to *(θ, φ, μ, σ)*. The second term refers to the Kullback-Leibler (KL) divergence between the approximate posterior distribution $q_\varphi(Z|X)$ that the encoder network maps the original data space into, as well as the pre-specified prior. In case of continuous latent variables, the prior is typically assumed to be Gaussian N(0, 1).

We propose that a VAE can outperform a normal AE for our context. However, VAE is difficult to optimize because the sampling in latent representation is not differentiable. Once we have the VAE network architecture defined, we re-parameterize it to make the VAE model trainable. Re-parameterization can stretch the encoded standard deviation with additional random noise, which propagates the reconstruction error to the encoder. Therefore, the tuning of parameters of VAE were optimized by using an Adam optimizer [185].

We constructed the VAE encoder with two stacked convolutional layers. Each convolutional layer is especially designed for FC input data. The details of this design are elaborated later. In addition, the decoder is constructed by two stacked deconvolution layers [186] [187]. The present work is the first to apply VAE for extracting representative features from HCP FC data [188] [189] in the source domain and then use the pre-trained model in the target domain classification task using ABIDE.

3.2.4. VAE pre-training and data generation model

As mentioned before, we use the VAE for pre-training and data oversampling. The idea is to let the VAE model learn latent representations of the data in an unsupervised way. In doing so, it is imperative to provide training examples from both classes. We used training data identified from ABIDE. As mentioned before, this is a total of 705 subjects, consisting of 277 ASD subjects and 428 controls subjects. However, just using the ABIDE training data in pre-training and then proceeding to classification by a CNN model will be the traditional approach devoid of any transfer learning. In order to bring in the component of transfer learning, we considered additional healthy control training examples from HCP database, consisting of 1097 control subjects. Therefore, in total, we had 1525 (1097 from HCP plus 428 from ABIDE) control subjects and 277 (from ABIDE) ASD subjects in training. This leads to class imbalance of 1248 samples (1525-277). In order to

alleviate this class imbalance, we let the VAE learn latent representations of ASD data and generate an additional 1248 synthetic ASD FC matrices. The procedure, illustrated in Fig. 7, solves the problem of class imbalance while simultaneously enabling transfer learning from the HCP data. This approach could in-principle be used to add any number of healthy control samples that are more widely available publicly and hence is a key component and innovation in our work.

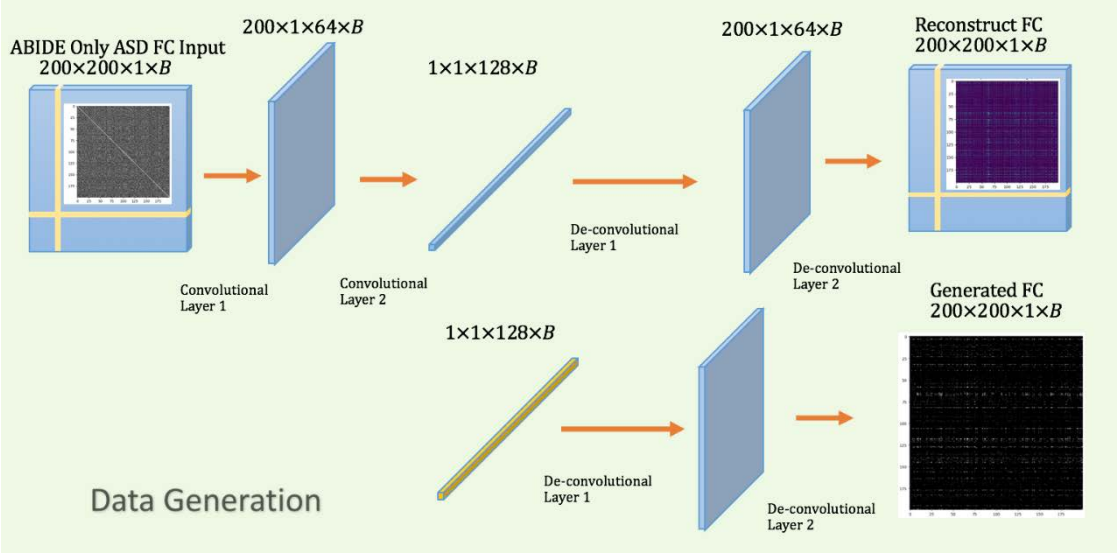

Figure 7 *Architecture of the VAE-CNN data argumentation model is illustrated. This approach was adopted to solve the class imbalance between controls and ASD arising from transfer learning of control samples from the HCP data. No dropout layer was applied in the model*

3.2.5. CNN classification model

Our FC-adaptive CNN classifier (i.e. a CNN model that has been adapted to suit FC features) consists of two 1D convolutional layers with the first layer for row connectivity convolution and the second layer for column connectivity convolution [154] [190] [191]. Each convolutional layer uses the ReLU activation function [192], and the output layer uses the Softmax function to obtain one-hot final result (a vector of each prediction with the predicted class as one

and others as zeros). In addition, we applied dropout layers after all convolutional and fully connected layers to increase the sparsity of the deep neural network [193] [194].

The architecture of the FC-adaptive CNN is introduced in Fig.5. To apply regularization, it includes multiple well documented layers. Similar to the E2N layer in the BrainNetCNN model [154], the kernel size of the first row convolutional layer is 200×1. The kernel size of the second column convolutional layer is 1×200. No pooling layer is included in FC adaptive CNN. The first one has a weight size of (128,64) and bias size of (64,1). The second layer has a weight size of (64, 32) and bias size of (32,1). The third layer has a weights size of (32,2) and bias size of (2,1). Drop-out layers are applied after each convolutional layer. Each layer randomly shut down 0.7 proportion of neurons in training iterations. The Softmax layer can calculate the probability array into a one-hot labels array. The predicted labels were then compared with the ground truth labels to measure the classification accuracy. The final cost includes the training loss and L1 regularization term.

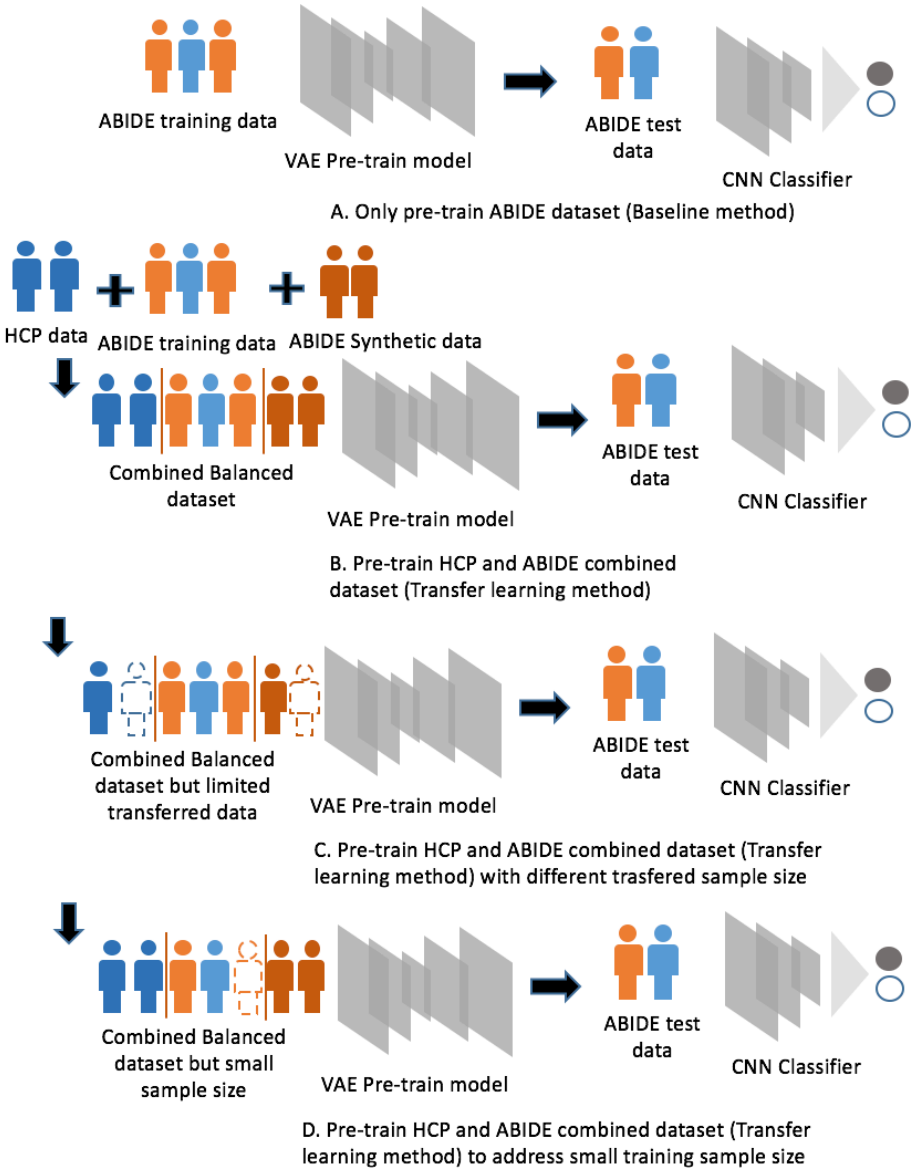

Figure 8. *Illustration of two strategies for ASD classification:* ***(A)*** *For the baseline method, we input ABIDE training and test datasets to the VAE-CNN framework.* ***(B)*** *For the transfer learning approach, we input the HCP and ABIDE balanced datasets into the VAE pre-train model and tested the classifier on the ABIDE test dataset.* ***(C)*** *To investigate the effect of transferred data size, we used half of combined balanced dataset of the previous approach and repeated the procedure.* ***(D)*** *To investigate the effect of small sample size, we used half the original training data in the transfer learning approach and repeated the procedure. Color code: light blue for ABIDE healthy subjects, light orange for ABIDE ASD subjects, dark blue for HCP healthy subjects and dark orange for ABIDE synthetic ASD subjects.*

3.2.6. Transfer Learning performance estimation

To measure the performance of our TL framework, we test the ASD classifier under four approaches (Fig.8). (1) As a baseline method (Fig.8A), we followed an autoencoder based DL approach similar to prior research [195] [196] [197], with only the ABIDE dataset as source domain data being used in training the VAE pre-train model. The training dataset included both heathy control (light blue, N=478) and ASD subjects (light Orange, N=277) from the ABIDE dataset. Then we used the CNN classifier to test the model on ABIDE test dataset (N=190). (2) To more specifically investigate the effect of TL (Fig.8B), we used the combined HCP and ABIDE dataset in VAE training (N=1525) for each of ASD and control groups, with the ASD group augmented by the VAE generative model as explained before). For each batch of HCP and ABIDE, we input HCP data first, which includes only heathy controls (dark blue), followed by the ABIDE dataset. (3) To measure the effect of the size of the source domain dataset on TL-based prediction (Fig.8C), we input only half of the source domain data (N=548 instead of 1097) and corresponding synthetic data in the pre-train model. (4) To test the capability of the TL model to potentially compensate for smaller sample size of the target domain data (Fig.8D), we input only half of the training samples in ABIDE (N=350 instead of 705). The estimation method was applied to both site-matched and site-mismatched datasets to measure the robustness of the model to multi-site effect.

Classifier performance was estimated by three diagnostic metrics: classification accuracy, balanced accuracy, and area under receiver operating characteristic curve (AUC). Classification accuracy is measured as the percentage of labels correctly predicted by the classifier on unseen test data. Given that ASD and control classes are unbalanced within both training and testing

datasets, we calculated balanced accuracy as well, which is measured by normalizing true positive and true negative predictions by the number of positive and negative samples (TPR and TNR, respectively) and then calculating the mean of the two values [198]:

$$\textit{Balanced Accuracy} = (TPR+TNR)/2$$

In addition, ROC (receiver operating characteristic) reflects the diagnostic ability of a binary classifier system when its cutoff varies. AUC is the area under the ROC curve. This was implemented using DL libraries Tensorflow [199] and PyTorch [200].

3.2.7. Layer-Wise Relevance Propagation (LRP) algorithm

After training the deep network, we applied the LRP algorithm to investigate input features important for classification [201]. The core idea of LRP is to trace back the contribution to the final output neuron layer by layer. The basic principle of LRP is that the total relevance of an output node to a class is conserved by each layer. Each of the nodes in the l^{th} layer that contribute to the activation of a node *j* in the next $(l+1)^{th}$ layer a part of the relevance in node *j*: R_{l+1}^{j}. In other words, the relevance of a node in a layer is the sum of all relevance the nodes in previous layer contributed to:

$$\sum_{i} R_{l,l+1}^{i\to j} = R_{l+1}^{j}$$

There are different ways in which the relevance can be distributed over the input neuron *i*, and different rules for how to distribute the relevance have been proposed [202]. In this paper, for

better visualization, we apply the ω^2 rule [203] to define the propagation of relevance from the $(l+1)^{th}$ layer to the l^{th} layer. The ω^2 rule formula is given by [204]:

$$R_i = \sum_j \frac{\omega_{ij}^2}{\sum_i \omega_{ij}^2} R_j$$

Here, ω_{ij} is the vector of weight parameters that connects neuron x_i with x_j. R_i and R_j represent the relevance scores of neurons x_i and x_i. The propagation rule consists of redistributing relevance according to the square magnitude of the weights and pooling relevance across all neurons j. This rule is also valid for $R_j = 0$, where the actual point $\{x_i\}$is already a root, and for which no relevance needs to be propagated. For further details, please refer to similar implementations in earlier work [205] and [162]. A PyTorch implementation of the LRP algorithm was developed for the current work and is available on GitHub (https://github.com/moboehle/Pytorch-LRP).

3.2.8. Feature identification

The input features were a 200×200 FC adjacency matrix, and the output was a one-hot prediction array. We calculated the relevance values that propagate backward from the output of one label to input features layer by layer. Then, because we are interested in FC features that have best predictive ability, the output of ASD prediction was traced back to the input layer by the LRP algorithm.

The procedure to identify FC biomarkers is illustrated in Fig. 9. The final heat map obtained from the LRP algorithm consisted of relevance scores of each feature in the input for a given

subject. The mean relevance values of each feature across subjects (within a given group) were calculated as a heat map representing the group. Then we used a permutation test to identify the statistical significance of the relevance values. We generated 1000 input FC maps from randomly selecting FC data and shuffling the subject labels. These surrogate data also went through the exact same pipeline as the original data and the LRP algorithm gave surrogate heat maps in each case. Thus, for each feature, we obtained a surrogate distribution of relevance values. By comparing the true relevance value with this surrogate distribution, we obtained a p-value for each input feature. The p-value of feature A is calculated as:

$$P_A = \frac{Num_{greatTrue}}{Num_{permutation}} \times 100\%$$

where $Num_{greatTrue}$ refers to the number of permutation values greater than true value, $Num_{permutation}$ refers to the permutation number that is equal to 1000. The significant features were determined in the heat maps based on an FDR-corrected p-value threshold of 0.05. To further check whether the identified features had stronger or weaker connectivity in ASD as compared to controls, we used two sample t-tests. These were then visualized using the BrainNet Viewer toolbox [206] in MATLAB.

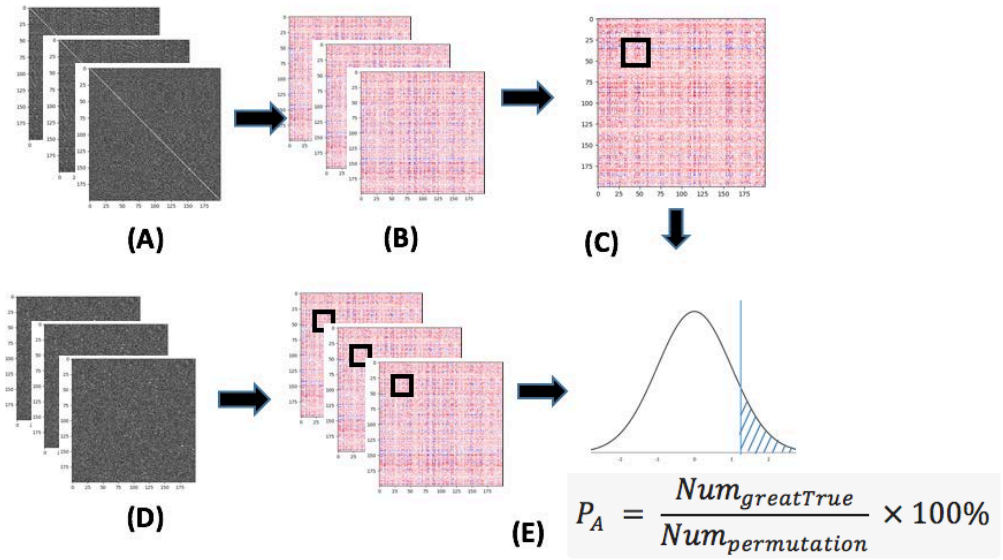

Figure 9. *Identifying FC features with high predictive ability using the LRP algorithm.**(A)** FC matrices from test subjects were input to the VAE-CNN transfer learning model. **(B)** The FC heat map for each subject was constructed using the LRP algorithm. **(C)** The true relevance score was calculated as the mean relevance across all subjects within a group. **(D)** Using 1000 randomizations, surrogate FC maps were created by random shuffling, and their heat maps were created using the same procedure used on the test dataset. **(E)** The final p-value for each FC feature was calculated from the true values and their corresponding surrogate distributions.*

3.3. Results and Discussion

Fig. 10 compares the classification performance between baseline classification and the proposed TL framework. To address the multi-site issue, we compared accuracies between site-matched and site-mismatched test datasets. In the site-matched case, the training and test data had subjects drawn proportionately from all sites. In the site-mismatched case, the training data were drawn from a few sites and the test data were from entirely different sites. In both cases, we found that the TL approach achieved higher accuracy (by about 7%) versus baseline classification. Accuracy with the site-mismatched dataset was also lower than with the site-matched dataset.

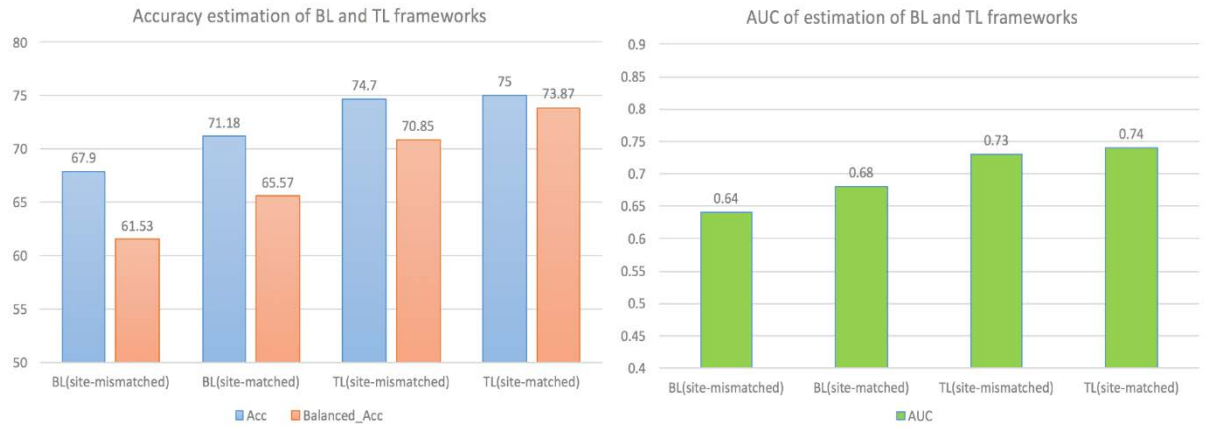

Figure 10. *Results obtained from ABIDE test data with and without transfer learning for both site-matched and site-mismatched cases (left: classification accuracies and right: total area underneath ROC curve (AUC)). Training and testing data were from different sites in the site-mismatched case, and drawn proportionally in the site-matched case. We found higher accuracy, balanced accuracy and AUC with transfer learning (TL) compared to baseline (BL) for all cases.*

The AUC in Fig.10 shows that transfer learning using the combined VAE-CNN classification model can achieve 0.73 AUC on site-mismatched dataset and 0.74 AUC on site-matched dataset, outperforming those obtained without TL. While classification with site-mismatched data deteriorated performance in the traditional baseline model in comparison to site-matched data (as expected [115]), TL helped the site-mismatched data catch up with the site-matched case by inclusion of more HCP data in the TL model. Thus, even though HCP data are from single site, it helps the model learn the variability in the healthy control population from a source other than ABIDE. This improves the generalizability (and hence, the performance) of the classifier. The performance improvement on unseen test data could in-principle be further boosted by exposing the model to more variability in both or either classes.

While it is encouraging to see the superior performance of the TL model, a critical question is whether the performance is dependent on the sample size in the HCP source domain. In order to answer this question, we repeated the entire analysis pipeline using only half of the original HCP

samples. We found (Fig. 11 and Fig. 12) improved performance with increased sample size of the source domain HCP data for both site-matched and site-mismatched cases. This is encouraging because it indicates that there is further scope to improve performance in the future by adding more healthy control data in the source domain using other public databases that are freely available.

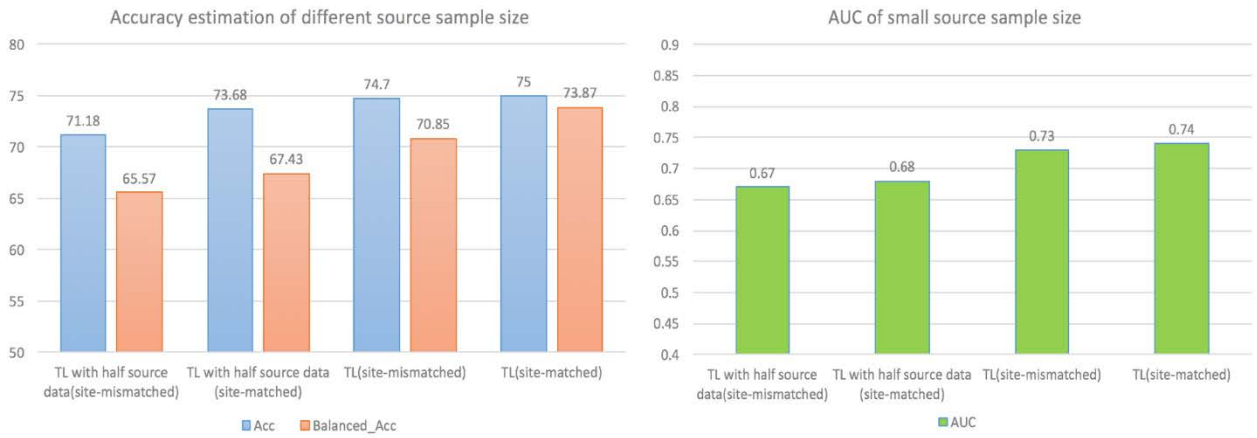

Figure 11. *Results obtained by using only half the samples in the HCP source domain and its comparison with those obtained by using the full HCP sample in the source domain (left: classification accuracies and right: total area underneath ROC curve (AUC)). The classification accuracy is shown in blue, the balanced accuracy as red and AUC as green.*

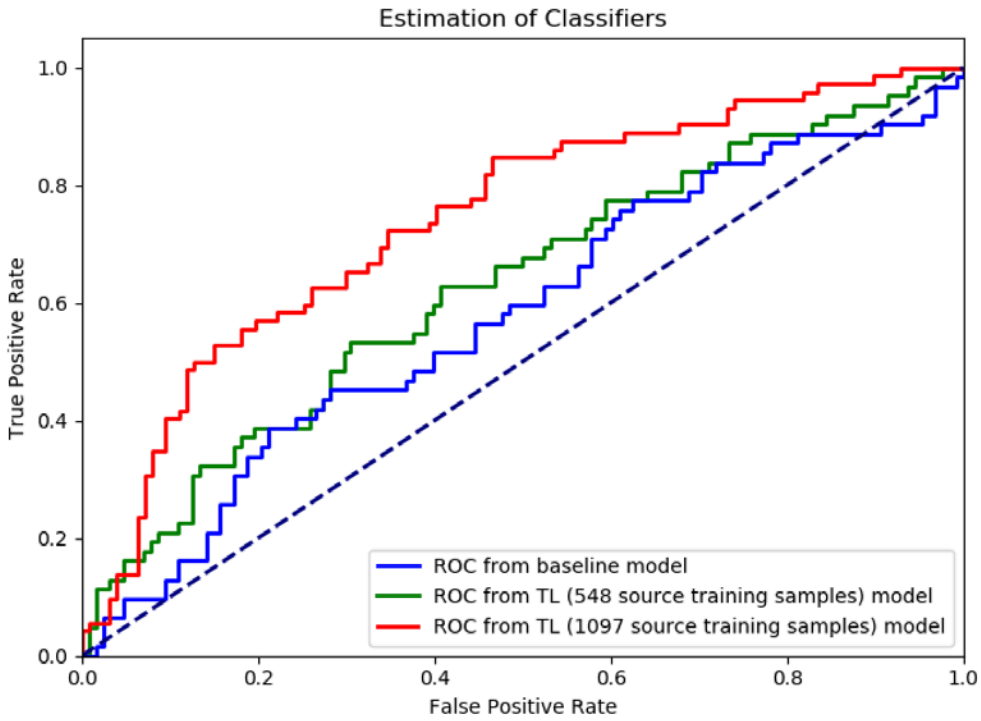

Figure 12. *ROCs obtained from site-mismatched test data with and without transfer learning approaches. In the transfer learning approach, the ROC curve achieved better shape when using the full HCP sample as the source domain compared to using only half of the HCP samples as the source domain*

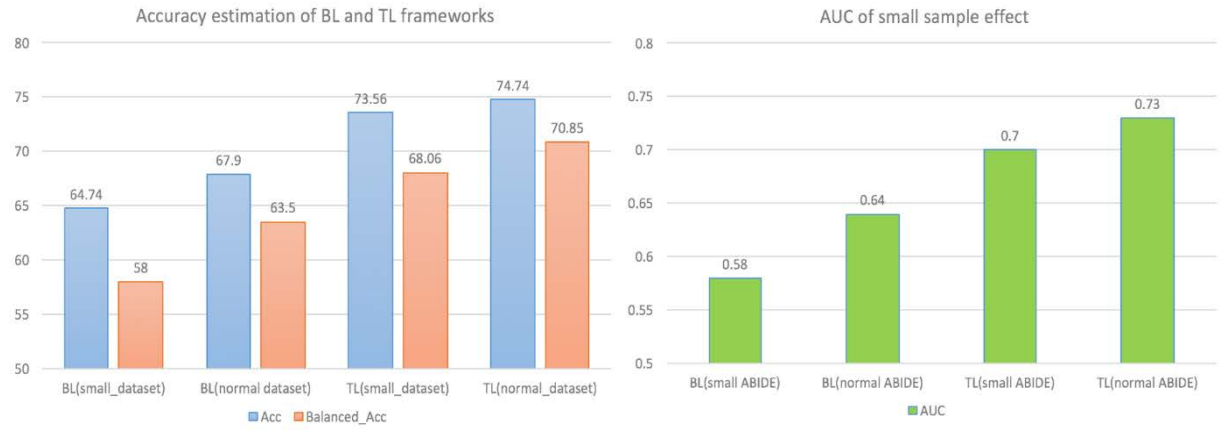

Figure 13. *Results obtained by using only half the samples in the ABIDE target domain data and its comparison with those obtained by using the full ABIDE sample in the target domain (left: classification accuracies and right: total area underneath ROC curve (AUC)). The classification accuracy is shown in blue, the balanced accuracy as red and AUC as green.*

To investigate the effect of the size of the target domain dataset on performance, we used only half of the ABIDE samples (N=350) as training data. The bar chart in Fig.13 illustrates that AUC from the smaller target domain training sample size is lesser than that from the full training sample size (N=705), because small training sample sizes cannot fully represent the overall data distribution in the test dataset. However, with TL in the VAE-CNN model, the difference in balanced classification accuracy between small and normal datasets was reduced from around 6% to 3%. The same result is also reflected in the AUC plot.

Finally, the LRP algorithm identified FC features that were important for classification in our model. Among connections identified as having significant relevance ($p<0.05$, FDR corrected), those with hyper-connectivity in the ASD group as compared to the control group are shown in Fig.12 and those with hypo-connectivity are shown in Fig.13. Overall, there were more connections with hypo-connectivity in ASD compared to hyper-connectivity. This is not surprising given that hypo-connectivity has been an established hallmark of ASD [207] [208] [209] [210]. Significant hyper-connectivity was mainly observed in local pathways within the frontal lobe (Fig.14), while hypo-connectivity was observed in a larger number of inter-hemisphere and inter-lobular connections in ASD (Fig.15). This is in agreement with the hypothesis [210] that ASD is characterized by local hyper-connectivity and distant hypo-connectivity. In Fig.15, many paths connect superior frontal regions as well as temporal gyrus and lingual gyrus. This anterior-posterior FC disruption is a commonly observed pattern in rs-fMRI studies of ASD [211]. The medial temporal lobe network, which incorporates the insula, and shows reduced inter-network connectivity, has also been found through ICA analysis of FC data [212].

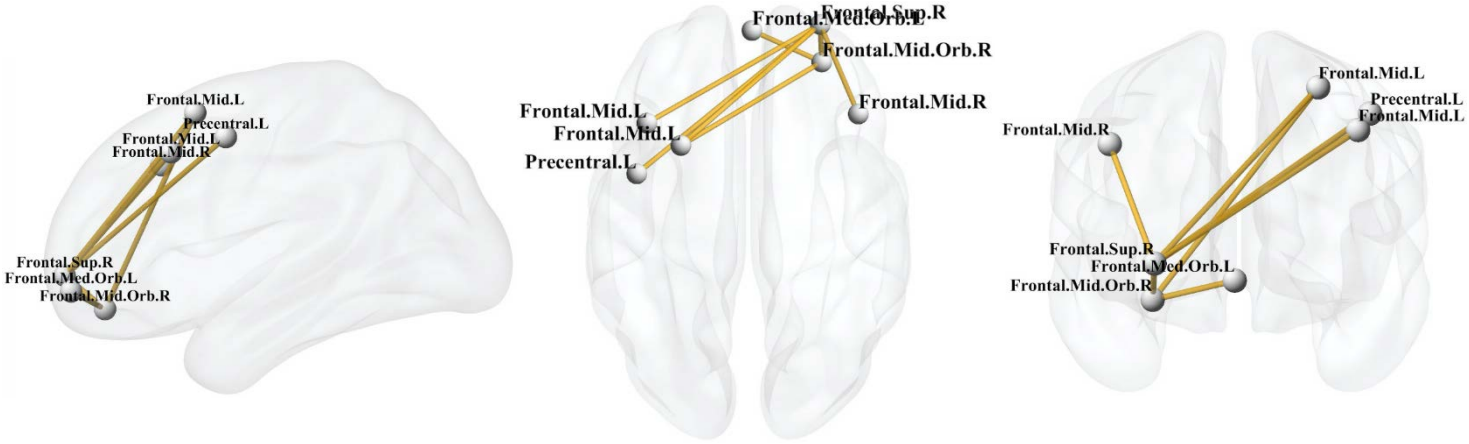

Fig 14. *Connections important for discriminating between controls and ASD (hyper-connectivity) obtained from the heat map generated by the LRP-based VAE-CNN transfer-learning classifier. Orange lines refer to hyper-connectivity in ASD versus controls.*

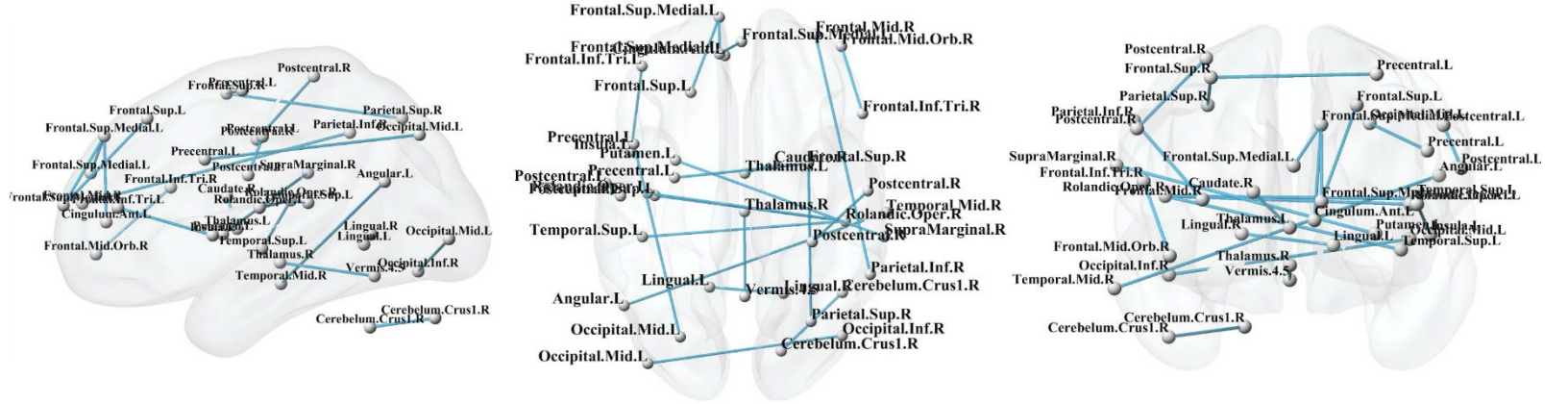

Fig 15. *Connections important for discriminating between controls and ASD (hypo-connectivity) obtained from the heat map generated by the LRP-based VAE-CNN transfer-learning classifier. The blue lines refer to hypo-connectivity in ASD versus controls.*

Chapter 4

VAE deep learning model with domain adaptation and harmonization for diagnostic classification from multi-site neuroimaging data

Abstract

In large public multi-site fMRI datasets, the sample characteristics, data acquisition methods and MRI scanner models vary across sites and datasets. This non-neural variability obscures neural differences between groups, leading to poor machine learning based diagnostic classification of mental disorders. This could be potentially addressed by domain adaptation, which aims to improve classification performance in a given target domain by utilizing the knowledge learned from a different source domain by making data distributions of the two domains as similar as possible. In order to demonstrate the utility of domain adaptation for multi-site fMRI data, we developed a variational autoencoder – maximum mean discrepancy (VAE-MMD) deep learning model for three-way diagnostic classification of Autism, Asperger's syndrome and typically developing controls. We chose ABIDE II (Autism Brain Imaging Data Exchange) dataset as the target domain and ABIDE I as the source domain. We show that domain adaptation from ABIDE I to ABIDE II provides superior test accuracy of ABIDE II as compared to using just ABIDE II for classification. Further, augmenting the source domain with additional healthy control subjects from Healthy Brain Network (HBN) and Amsterdam Open MRI Collection (AOMIC) datasets enabled transfer learning and further improved performance on ABIDE II classification. Finally, comparison with statistical data harmonization techniques such as ComBat revealed that deep learning models such as VAE-MMD, when used in combination with statistical methods, can

provide incrementally better performance. We openly share our data and model so that the possibility of further improvement of the model, by utilizing the ever-increasing amount of healthy control fMRI data in the public domain, can be explored by the neuroimaging community.

Keywords: *functional connectivity, autism spectrum disorders, domain adaptation, variational autoencoder, machine learning prediction*

4.1. Introduction

Deep learning models outperform traditional machine learning methods in identifying individuals with psychiatric disorders, including autism [213] [69]. However, they require larger sample sizes to avoid overfitting [214]. Large public databases such as ABIDE (Autism Brain Imaging Data Exchange) have aided deep learning models in this endeavor. However, such large public databases have been assembled post-hoc, and hence contain different sources of non-neural variability such as different sites using different scanners and protocols. Normally, the samples from different scanners or acquisition protocols do not follow the same distribution in most cases [44]. If the test data and training data are drawn from different independent distributions, the performance of deep learning as well as traditional machine learning models will be degraded [213] [115]. To address this, we propose domain adaptation, which aims to improve the classification performance in a target domain by utilizing the knowledge learned from the source domain by making the distributions of data in source and target domains as similar as possible [215] [216] [217].

Although, multiple previous frameworks have been developed to exploit commonalities between different data domains to achieve domain adaptation in various areas [218] [219] [220] [221] [105] [222], limited number of end-to-end deep learning models incorporating domain

adaptation have been developed for neuroimaging data (wherein the target system is completed by one deep learning model instead of a sequence of intermediate steps). For example, Li and colleagues [216] proposed a domain adaptation framework on federated datasets across different sites of the ABIDE dataset. Similarly, Zhou and colleagues [217] formulated the DawfMRI framework, which revealed additional insights into psychological similarity among the OpenfMRI project databases. Both of them aligned different domains of data into one common embedding space followed by biomarker identification. But it is achieved by training each local model individually and then integrating them together by an ensemble strategy. Since this is not implemented as a single deep learning model, complexity increases, ease of use decreases and it becomes more difficult to train and optimize. Moreover, Wachinger et al. [223] validated that a supervised domain adaptation framework yields better results than the simple union of source and target training sets in Alzheimer's disease diagnosis. Nevertheless, their methods were developed by an instance weighting strategy optimization combined with a shallow machine learning algorithm instead of an end-to-end domain adaptation deep learning model [224]. With deep learning models have been validated to learn features from neuroimaging data in multi-site studies [225] [226] [227], it is obviously a better choice to use in a domain adaptation framework. Therefore, a single and integrated domain adaptation deep learning model has potential and is preferable, but has not yet been tested and validated for classifying psychiatric disorders from neuroimaging data.

Existing domain adaptation approaches applied in neuroimaging based diagnostic classification primarily employ supervised learning. For example, a previous study [228] proposed a robust domain transfer support vector machine (DTSVM) to classify mild cognitive impairment (MCI) by using the labeled Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

Another deep domain adaptation method [229] utilized the supervised domain adaptation (SDA) method on the pre-trained VGG network, which is a popular deep convolutional network developed by Simonyan et al [230], and used labeled MRI data to fine-tune the model for Alzheimer's disease prediction. Nevertheless, as we know, developing prediction models on medical data is marred with the complex labeling process that is not always accurate [231]. This is because diagnosis of psychiatric disorders is based on behavior and not objective biomarkers. This can make the labels less accurate for marginal cases and stratification of individuals in spectrum disorders. Since label scarcity is a common challenge across medical imaging studies, unsupervised domain adaptation (UDA) has gained importance recently [232]. UDA is also a solution to address potential inaccuracies in labels [214] and to also increase the statistical power of analysis by adding more unlabeled data [233]. By combining the advantages of supervised and unsupervised learning DA methods, semi-supervised approaches for domain adaptation improve accuracy of discriminative prediction in the UDA scenario [232]. This class of methods only require limited quantity of labeled data or no labeled data from the target domain [234]. However, literature on this semi-supervised learning approach in psychiatric disorder classification is scarce. To explore and validate this approach in the current study, we used unlabeled data during training in the target domain using a semi-supervised approach to achieve domain adaptation.

Some previous studies have developed semi-supervised domain adaptation methods and tested them on deep learning benchmark data [235] [236]. Among them, variational autoencoder (VAE) [237] has been demonstrated to be robust against high-dimensional input data and is able to flexibly learn various distributions. By using this advantage of VAE in learning the features, [238] variational fair autoencoder (VFA) was further developed to learn the features that are invariant to

noisy nuisance factors, but retained useful information as much as possible. Developed from VFA, the domain invariant variation autoencoder (DIVA) [239] generated cell image data that matched the ground truth factors of interest but from a previously unseen domain. The VAE-based domain adaptation model is capable of extracting class features that are invariant to different domains. One semi-supervised learning framework, based on VAE with recurrent inference model, was used to construct the domain mixed feature representation from clinical care data [240]. In [238]. In addition to VAE, the discrepancy between marginal posterior distributions of the data can be minimized by a maximum mean discrepancy (MMD) layer [241]. Specifically, during feature extraction and training, MMD regularization term and VAE were jointly optimized by Gretton et al. Given these advantages of VAE and MMD in semi-supervised domain adaptation, we set out to test the validity of this model on high dimensional functional MRI (fMRI) data [242] in the context of domain adaptation for diagnostic classification from multi-site fMRI data. There is no other fMRI study that used the same approach before.

In this study, we propose to use variational and adversarial classification frameworks for domain adaptation by training labelled data in the source domain and unlabeled data in the target domain. Variational inference model was used to learn the invariant representations across information from different sites of the ABIDE dataset, while retaining the discriminative information in the classification task. We applied a model based on VAE, which can naturally encourage separation between latent feature representations and domain variables. However, some dependencies can still remain if the labels of data points are correlated with the domain variable, which can ‘leak’ some of the domain information into the latent feature representation, resulting in dependency. Thus, our model uses a “maximum mean discrepancy” [241] regularization term, which is a measurement of divergence between two distributions, to penalize the distances between

the latent probability distribution across source and target domains. During the adversarial training procedure, the domain ‘confusion’ is maximized to ensure that the features are domain invariant, and the classification of ASD is also optimized.

To augment domain adaptation and improve the generalizability of the classifier, we include more data from Healthy Brain Network (HBN) [243] and the Amsterdam Open MRI Collection (AOMIC) [244] datasets in the source domain. More specifically, HBN provides the research community with a large-scale dataset of over 10,000 healthy children through an open data sharing mode. AOMIC contains large-scale resting-state fMRI data from healthy individuals collected at the University of Amsterdam over the past decade. From the demographic information in Table.1, the age range of HBN and AOMIC are close to ABIDE I and ABIDE II. One reason we included these two databases in the domain adaptation model was to increase the variety of data distribution and enhance the generalizability of the model.

We compare and contrast the proposed method with ComBat harmonization [245], which is a statistical technique used to reduce divergence of data distributions from multi-site MRI data. This is considered a current gold standard and hence, we compared and combined ComBat harmonization method with the proposed deep learning approach in this study [245]. Various approaches have been developed to remove the undesirable inter-subject or inter-site factors and make the data more easily comparable [246] [247]. Among those methods, ComBat harmonization has been applied on neural imaging data across scanners and sites [248]. Specifically, ComBat harmonization [249] focuses on dealing with the variability of parameters’ distributions so that they can be pooled together. The source of the variation, the so-called “batch effect” [245], is eliminated primarily based on an empirical Bayes framework [249] [248]. ComBat was also proposed to correct for site effects in functional measurements from multi-site fMRI data [250].

We applied ComBat harmonization to the input data in this study as one of the methods to reduce the domain shift.

Finally, identification of imaging features that are important for diagnostic classification is crucial for ASD biomarker discovery and diagnosis [251]. The interpretation of the correlation between domain adaptation and selected features is still a challenge in multiple studies [252] [253] [254] [255]. Especially in this study, imaging features in the VAE based model are difficult to trace back from the output layer to the input layer because of the continuous Gaussian latent variables in the latent space [256]. We thus propose a statistical method to identify such imaging features.

Based on the information presented above, we summarize four major aspects of the proposed framework:

1. We use a VAE-MMD model for domain adaptation in multi-site fMRI data for predicting the diagnostic labels from fMRI functional connectivity (FC) data. We demonstrate that domain adaptation from the first release of the ABIDE dataset (ABIDE I) to the second release (ABIDE II) will improve classification performance of ABIDE II as compared to performing classification solely on ABIDE II.
2. We compare and contrast statistical (ComBat) with deep learning (VAE-MMD) approaches for domain adaptation.
3. We test whether additional data in the source domain, specifically healthy control data, will augment domain adaptation and improve the generalizability of the classifier in achieving better accuracy in the target domain of ABIDE II. Given the large amount of healthy control data available in the public domain, this approach could potentially be used to substantially

improve diagnostic classification in relatively smaller public datasets obtained from individuals with mental disorders (such as ASD)

4. We extract and identify imaging features diagnostically important for ASD prediction across different fMRI data distributions.

4.2. Methods

4.2.1. The fundamental algorithm of a neural network

4.2.1.1. Multi-layer perceptron (MLP)

Deep learning algorithms have complex mathematical structures with several processing layers that can extract data features into various abstraction layers. The building block of DNN, the MLP [257], is a typical type of layer in feed-forward networks in which each node is connected to all the nodes in the next layer. Within each node in MLP, the input values are combined with weights and bias, and then summed up before being passed to an activation function. The most used activation functions include sigmoid, tangent hyperbolic (tanh) [258] and rectified linear unit (ReLU) [259]. The output *z* of a node in an MLP layer can be calculated as:

$$z = \sigma(\sum_{i=1}^{m} w_i x_i + b)$$

where *m* refers to the number of nodes in the current layer, *w* corresponds to the weights of all connections between the current node and nodes in the previous layer, *b* corresponds to bias and σ corresponds to a non-linear activation function.

4.2.1.2. Training an MLP

The weights of biases of the MLP are trainable parameters, which are optimized during the training process. Normally, those parameters are initialized with random variables close to zero. After the forward computation of the MLP, the loss function can be defined as the mean squared error (MSE) in single class scenarios and cross-entropy in multi-class scenarios. In the training procedure, the MLP weights can be learned by training with a basic error back-propagation technique for the loss function. Back-propagation is based upon an optimization algorithm using stochastic gradient descent (Bottou, 2012) with a pre-defined learning rate. During each round of computation, the values of the network parameters can be optimized by computing the gradient of the loss function with respect to each of them using the chain rule.

The input data of MLP always separates into groups, and each group of samples is called a batch. The number of samples in the input group is referred as batch size. After all the data are trained, the procedure repeats a certain number of times called an epoch number. Different from batch, an epoch indicates one iteration of the entire training dataset the ML model has completed. The number of entire iterations is named as epoch number. Except the trainable parameters optimized during training procedure, pre-defined parameters such as batch size, epoch number or learning rate are fixed during training and are referred to as hyper-parameters.

4.2.1.3 Overfitting and regularization

Overfitting occurs when a well-trained MLP fits accurately to the training data but performs poorly with the unseen test data. Especially in neuroimaging, the training sample size is limited

[260], which is problematic for generalizing the findings to a clinical setting. Two straightforward ways to address overfitting are simplifying the model and increasing the training sample size. In addition, overfitting can be addressed by adding regularization to the objective function. Those modifications such as the well-known L1/L2 terms (Ridge and Lasso Regression) cause the model to be simpler during optimization but enhances the generalizability on unseen data [261].

4.2.2. Baseline techniques for ASD classification

Machine learning techniques such as SVM and MLP neural network performed well in previous ASD classification studies [262] [263] [264] [265]. To estimate the performance of the proposed domain adaptation approach, we designated traditional SVM and MLP as baseline approaches in this study. Specifically, SVM used a polynomial kernel, and the hyper-parameter *C* was set to 100. In our implementation, the architecture of the baseline MLP method was the same as that of the VAE used in domain adaptation. The architecture has two layers, the first layer contains 200 nodes, and the second layer contains 500 nodes.

4.2.3. Participants and Data

Our aim was to test the utility of our domain adaptation model on fMRI data. We used ABIDE resting-state fMRI data for this purpose [266]. We used ABIDE I [68] (released in August 2012) as the labeled supervised dataset and ABIDE II [69] (released in June 2016) as the unlabeled semi-supervised dataset. To investigate the domain adaptation effect of our VAE-MMD model, we set ABIDE I as source domain, and ABIDE II as target domain dataset. There were a total of 998

subjects from 15 sites in the ABIDE I dataset and 623 subjects from 11 sites in the ABIDE II dataset.

Table 2: *ABIDE I data pooled from 15 different sites (and 18 cohorts, since some sites had more than one cohort), and ABIDE II from 11 sites. The acquisition sites include California Institute of Technology (CALTECH), Carnegie Mellon University (CMU), Kennedy Krieger Institute (KKI), University of Leuven (LEUVEN), Ludwig Maximilians University Munich (MAX), NYU Langone Medical Center (NYU), Olin Institute of Living at Hartford Hospital (OLIN), University of Pittsburgh School of Medicine (PITT), Social Brain Lab BCN NIC UMC Groningen and Netherlands Institute for Neurosciences (SBL), San Diego State University (SDSU, Trinity Centre for Health Sciences (TRINITY), University of California, Los Angles (UCLA),University of Michigan (UM), University of Utah School of Medicine (USM), Yale Child Study Center (YALE), Georgetown University (GU), Oregon Health and Science University (OHSU), Olin Neuropsychiatry Research Center (ONRC), Trinity Centre for Health Sciences (TCD), University of California Davis (UCD), University of Miami (UM). Across different acquired sites, the age and gender distributions change considerably. Both AOMIC and HBN data have had multiple releases.*

Database	Acquisition site	Subjects	Age Mean	Age Std	Male	Female
ABIDE I	CALTECH	32	26.79	9.6	25	7
	CMU	27	26.59	5.58	21	6
	KKI	55	10.1	1.31	42	13
	LEUVEN_1	29	22.59	3.49	29	0
	LEUVEN_2	35	14.16	1.4	27	8
	Max	57	26.16	11.98	50	7
	NYU	179	15.39	6.59	142	37
	OLIN	36	16.81	3.44	31	5
	PITT	57	18.9	6.82	49	8
	SBL	24	33	6.7	24	0
	SDSU	32	14.35	1.85	25	7
	TRINITY	42	16.84	3.63	42	0
	UCLA_1	73	13.16	2.38	63	10
	UCLA_2	26	12.49	1.5	24	2
	UM_1	107	13.43	2.87	83	24
	UM_2	35	15.96	3.27	33	2
	USM	100	22.14	7.67	100	0
	YALE	42	12.96	2.8	30	12
	TOTAL	**988**	**18.43**	**7.82**	**840**	**148**
ABIDE II	GU	104	10.68	1.62	69	35

	KKI	197	10.34	1.27	128	69
	NYU	27	6.78	1.07	24	3
	OHSU	91	10.88	1.99	56	35
	ONRC	43	23.33	3.85	31	12
	SDSU	23	13.91	3.85	20	3
	TCD	19	14.45	2.67	19	0
	UCD	32	14.78	1.83	24	8
	UCLA	32	10.7	2.36	26	6
	USM	32	21.37	7.74	27	5
	UMIA	23	9.8	2.02	17	6
	TOTAL	**623**	**13.37**	**2.75**	**441**	**182**
AOMIC	PIOP1	216	21.96	1.91	29	44
	PIOP2	226	21.96	1.79	96	129
	TOTAL	**442**	**21.96**	**1.85**	**125**	**173**
HBN	CBIC	287	10.75	3.73	188	99
	SI	345	11.13	3.82	195	150
	RU	753	9.92	0.42	501	252
	TOTAL	**1385**	**10.60**	**2.66**	**884**	**501**

ABIDE I fMRI data included 988 subjects from 15 different sites (and 18 cohorts, since some sites had more than one cohort). The number distribution of subjects across multiple sites is shown in **Table 2**. FMRI data were pre-processed using DPARSF [167]. This involved the removal of the first five volumes, slice timing correction, motion correction, co-registration to the standard MNI space, censoring of high motion volumes and regressing out nuisance variables (low frequency drifts, mean global signal, motion parameters, and white matter and cerebrospinal fluid signals). Voxel time series were temporally filtered with a 0.01–0.1 Hz bandpass filter. ABIDE II fMRI data included 623 subjects from 11 different sites. The pre-processing pipeline for this was identical to that used for ABIDE I (however, it was performed in CONN software) [267].

In order to test whether the generalizability of the model can be further improved by augmenting the size of the source domain by adding healthy control data, we used two additional datasets in this study: the AOMIC [244] and HBN datasets [243]. AOMIC (https://nilab-

uva.github.io/AOMIC.github.io/) data was organized from large-scale MRI projects at the University of Amsterdam to analyze individual differences in fMRI data. They publicly provided both raw and well-established preprocessed forms of three datasets: PIOP1 (Population Imaging of Psychology), PIOP2 and ID1000. Each of them have specific data acquisition protocols and participants. In this study, we used the raw PIOP1 (N = 216) and PIOP2 (N = 226) datasets instead of the preprocessed datasets. HBN (http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/) was acquired for detecting and characterizing pathologic processes in the developing human brain [243]. They have publicly shared a biobank comprised of data from 10,000 New York City area children and adolescents (ages 5–21) out of which only a small subset of subjects had an MRI. Here we use good quality MRI data from 1385 subjects in HBN. HBN data were collected from three sites: Citigroup Biomedical Imaging Center (CBIC), Staten Island (SI) and Rutgers University (RU). The demographic information of subjects in these two datasets is included in **Table 1**.

The pre-processing pipeline for all the datasets were identical (Figure 16). The use of additional source domains comprising of AOMIC and HBN datasets allowed us to test the assumption that when the size of the source domain or number of source domain subjects increases, it improves domain adaptation. Specifically, the HBN dataset contains data from children and may be relevant for domain adaptation when the target domain also contains data from children (such as ABIDE or ADHD-200), as in our case.

4.2.4. Feature extraction

We used the whole brain cc200 atlas [171] to reduce the dimensionality of the data. This atlas was generated spectral clustering of resting state fMRI data of healthy subjects, and hence, the ROIs in the atlas are said to be functionally homogeneous. Mean time series were extracted from

200 regions of interest (ROIs) of the atlas. We then estimated FC by computing the Pearson's correlation coefficient between each pair of time series. A vector of 19900 individual features per subject was constructed by reshaping the upper triangle of the 200×200 connectivity matrix minus the diagonal. Only the upper triangle was considered since FC is a non-directional metric.

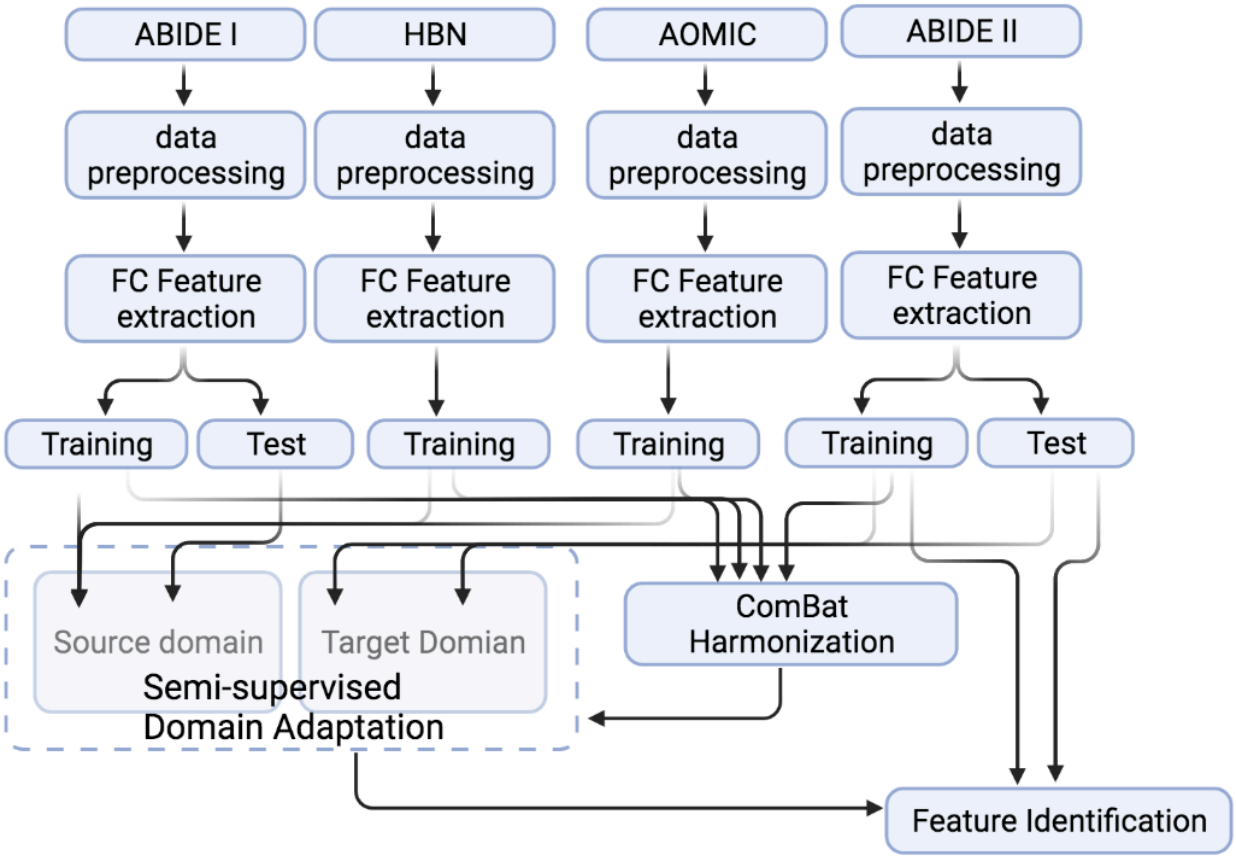

Figure 16. *A flowchart representation of the complete processing and analysis of multiple datasets. The fMRI data from ABIDE I, ABIDE II, HBN and AOMIC were subjected to identical data preprocessing and FC feature extraction. Source domain training and testing uses ABIDE I data while data from healthy subjects in AOMIC and HBN are used as additional training samples in the source domain to test the effect of source domain training sample size on domain adaptation performance. Target domain training and testing uses ABIDE II data. Only training datasets are harmonized by the statistical ComBat harmonization method and then input to the DA deep learning model, a procedure similar to the non-harmonized data. Note that only the target domain ABIDE II data is used for identification of features important for classification.*

4.2.5. Domain adaptation VAE-MMD model with semi-supervised learning

In order to realize domain adaptation, we apply the semi-supervised VAE model first proposed by Kingma et al. [237] with unsupervised learning. The model consists of a generative model $p_\theta(x|z,d)$ and an inference model $q_\phi(z|x,d)$, where z is the latent variable representation, x is the input data, and d is the domain variation we desire to remove. θ and ϕ are the trainable parameters of the generative model and inference model, respectively. For semi-supervised classification, our goal is to construct latent variable z, which has maximum information about the observed label y, while excluding the information about the nuisance domain variable d. It is achieved by adding an additional model in the generative model to correlate latent features to the classification task [238] . The schematic of the this model is shown in **Figure 17**, where the invariant feature in the first model M1 is referred to as z_1. M1 generates x as $x \sim p_\theta(x|z_1,d)$, and M2 generates domain invariant variable z_1 as $z_1 \sim p_\theta(z_1|z_2,y)$. y is a categorical variable that denotes the label of the data point x and z_2 encodes the variation on z_1 that is independent to y. Thus, for the N labeled data points and M data points without labels (i.e. unlabeled data), the objective function of VAE becomes:

$$\mathcal{F}_{VAE}(\phi,\theta;x_n,x_m,d_n,d_m,y_n) = \sum_{n=1}^{N}\mathcal{L}_s(\phi,\theta;x_n,d_n,y_n) + \sum_{m=1}^{M}\mathcal{L}_T(\phi,\theta;x_m,d_m)$$

$$+\alpha\sum_{n=1}^{N}\mathbb{E}_{q(z_{1n}|x_n,\ d_n)}[-logq_\phi(y_n|z_{1n})]$$

where the first and second terms denote the lost functions from the labeled and unlabeled data. Because the label predictive distribution $q_\phi(y|z_{1n})$ only contributes to the unlabeled data in the second term, we compensate this by adding a regularization term with weight coefficient α to

ensure that $q_\phi(y|z_{1n})$ is learned from both labeled and unlabeled data. Increasing α results in more purely discriminative learning in the generative model.

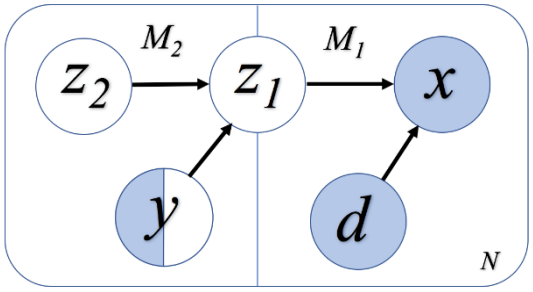

Figure 17: *A flowchart representation of the semi-supervised learning model. White variables refer to the variables without input information and blue variables refer to those with input information. Only some of the labels of y are known, and hence y is half white and half blue. We assuming the variables z_2 and y are independent to each other, while z_1 and d are independent to each other. Among them, z_2, y, and d are independent variables, and z_1 is dependent on z_2 and y.*

In the VAE inference model, we assume that variables z_1 and d are statistically independent to each other so that the marginal posterior distribution $q(z_1|d)$ is equal to zero. However, the independence relationship may fail because of the correlation between y and d. As discussed in the introduction section, we apply an additional MMD regularization term to penalize this situation.

In the MMD definition, the divergence between two distributions is calculated as the distances between mean embeddings of features [268]. Let k be a continuous, bounded, positive semi-definite kernel and H be the corresponding reproducing kernel Hilbert space [241], which are reduced by the feature mapping from X to H. The MMD of distributions $p_x(x)$ and $p_y(y)$ is defined as follows:

$$MMD(p_x, p_y) = \left\| \mathbb{E}_{x\sim p_x}[\varphi(x)] - \mathbb{E}_{y\sim p_y}\left[\varphi(y)\right] \right\|_H^2$$

In the VAE model, an additional MMD regularization term was applied to enforce the model to match the source and target domain marginal posterior distributions of latent variables $q(z_1|d = 0)$ and $(z_1|d = 1)$. So, the MMD term is determined as:

$$\ell_{MMD}\left(Z_{1,d=0}, Z_{1,d=1}\right) = \left\| \mathbb{E}_{\tilde{p}(x|d=0)}[\mathbb{E}_{q(z_1|x,d=0)}\left[\varphi(z_1)\right]] - \mathbb{E}_{\tilde{p}(x|d=1)}[\mathbb{E}_{q(z_1|x,d=1)}\left[\varphi(z_1)\right]] \right\|_H^2$$

Where d be the domain nuisance variable. Finally, adding the MMD penalty term into the lower bound of the aforementioned VAE, the proposed model becomes:

$$\mathcal{F}_{MMD-VAE}(\phi, \theta; x_n, x_m, s_n, s_m, y_n) = \mathcal{F}_{VAE}(\phi, \theta; x_n, x_m, s_n, s_m, y_n) - \beta\, \ell_{MMD}\left(Z_{1,s=0}, Z_{1,s=1}\right)$$

where β denotes the regularization coefficient in domain adaptation. Increasing β results in more domain confusion regularization compared to the classification loss. Both α and β are hyper-parameters that control the trade-off between classification loss and domain confusion loss, which are optimized through training and validation.

The datasets input to the semi-supervised learning model are illustrated in flowchart in **Figure 16** and the entire framework is illustrated in **Figure 18**. In the training model (#1 in **Figure 18**), we input both ABIDE I with labels and ABIDE II without labels as training datasets into the VAE-MMD model. The original t-distributed Stochastic Neighbor Embedding (t-SNE) figure and the corresponding t-SNE figures (**Figure 19**) after domain adaptation were generated in the beginning

and last iteration of this process. (t-SNE) [269] is a dimension reduction technique that allows us to visualize the group-wise separation of features in latent space and is used to visually assess the efficacy of domain adaptation (details described later in section 2.3.4). In the validation model (#2 in **Figure 18**), ABIDE I and ABIDE II validation datasets were used in the validation process for fine-tuning the hyperparameters α and β. At the end, ABIDE I and ABIDE II test datasets were used in testing the model to measure the model's performance (#3 in **Figure 18**). For better understanding of the model performance, we record accuracy from both training, validation and testing models.

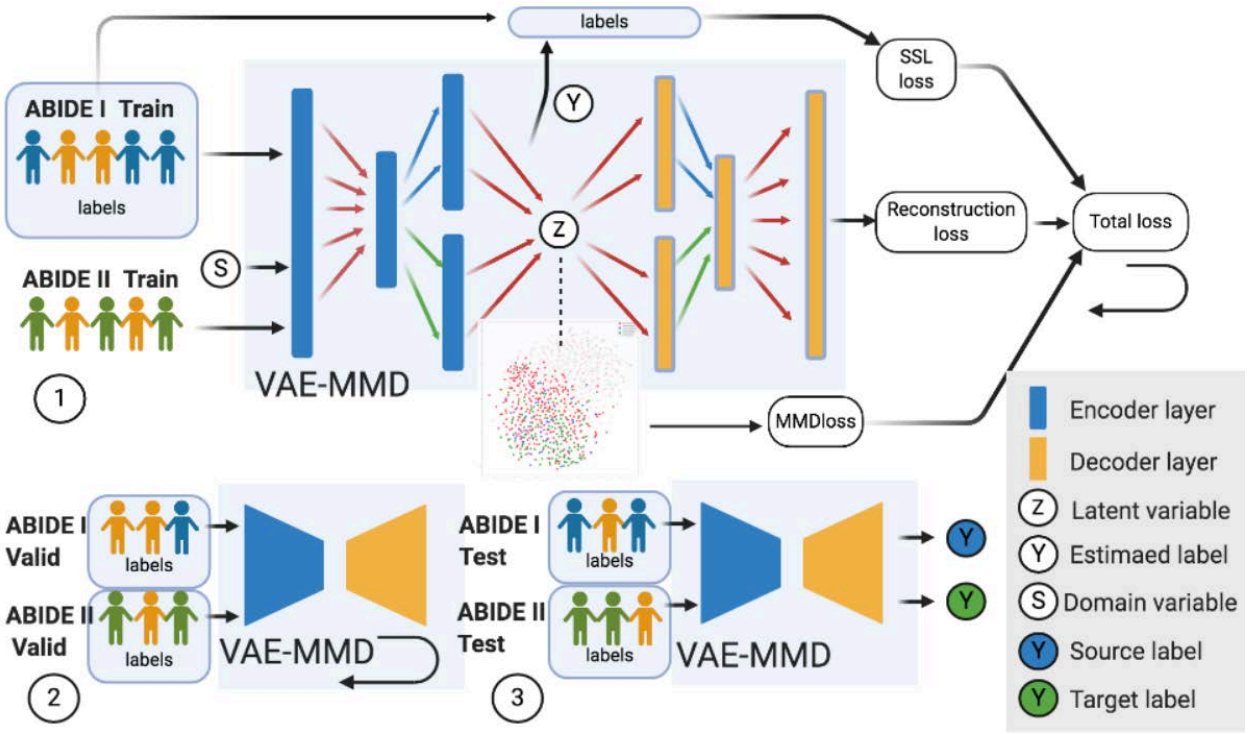

Figure 18: *Three major steps in the VAE-MMD model. (1) For training, we input both ABIDE I with labels and ABIDE II (without labels) training datasets into the VAE-MMD model. The original t-SNE figure and domain adapted t-SNE figure were generated in the beginning and last iteration of this process. The total loss was constructed by semi-supervised learning loss, reconstruction loss and MMD loss. (2) For validation, ABIDE I and ABIDE II validation datasets were used for fine-tuning the hyperparameters α and β. (3) For testing, ABIDE I and ABIDE II test datasets were used in testing the model to evaluate the model's performance. Subjects in orange represent healthy controls in ABIDE I and ABIDE II data, subjects in blue represent ASD subjects in ABIDE I and subjects in green represent ASD subjects in ABIDE II.*

4.2.6. Model setup

In deep neural networks, MLPs [270] work well on vector inputs while CNNs perform better on natural images by taking advantage of spatial dependencies. Since functional connectivity inputs can be vectors and are not natural images, we felt that the tradeoff tipped slightly in favor of MLPs. In the encoder, the first layer is constructed as the latent-feature discriminative model (M1) and the second layer is constructed as a generative semi-supervised model (M2) in a stacked architecture. M1 refers to the first layer and M2 to the second layer in the encoder of the #1 in Fig.18. The dimension of latent features in the first and second encoding layers were equal to 2000 and 1000, respectively. The learning rate was set to 0.0001. To reduce the likelihood of the gradients vanishing, each neural network layer used ReLU as activation function [271]. The epoch number was equal to 50 and the number of batches was 20. The code was constructed in Theano and Python software platforms.

We set ABIDE I as the source domain dataset, and ABIDE II as the target domain dataset. One of our goals was to reduce the non-neural differences in data characteristics between the two domains. The ABIDE I dataset was split into 673/157/158 subjects as training, validation and test datasets, respectively. The training and validation sets used labeled data in a cross-validation framework for fine-tuning hyperparameters. For the ABIDE II dataset, it was split into 371/126/126 subjects as training, validation and test datasets, respectively.

4.2.7. Transfer learning

Transfer learning (TL) [222] is a technique that applies knowledge learned from one domain and one task to another related domain and/or another task [265]. To improve generalizability,

address overfitting and increase sample size in the source domain, HBN and AOMIC data were included as additional source domain data. The labels of these two datasets are all healthy controls, which were used during training. The number of batches of HBN and AOMIC was equal to that of ABIDE dataset, so that they can be trained simultaneously. The divergences of these two datasets to the target domain data were also optimized during training, same as ABIDE I data in the source domain.

4.2.8. ComBat harmonization

We used the publicly available MATLAB toolbox [272] to achieve ComBat harmonization. We used default options recommended by the creators of the toolbox for its implementation. Finally, we separated harmonized data into training and testing datasets and input into the deep learning model, followed the same pipeline in training model as without ComBat harmonization (Fig.16).

4.2.9. Model estimation

Estimation of model performance happened at three levels. First, visualization of the separation of features in latent space with/without domain adaptation was realized using t-SNE plots. Next. Kullback–Leibler (KL) divergence was used to characterize the same analytically. Finally, model performance was characterized in terms of accuracy, etc. Details are described next.

To visualize the effectiveness of domain adaptation, we used t-SNE [269] as a dimension reduction technique in the latent space. t-SNE is particularly useful for the visualization of high dimensional datasets. Dimensionality reduction is important because if the latent variables have dimensions higher than three, we cannot visualize them. t-SNE is a method developed from the stochastic neighbor embedding (SNE) technique, converting the high-dimensional Euclidean

distances between data points into conditional probabilities that represent similarities between data points. The clustering of data points is based on the similarities across data points. In comparison with SNE, the cost function used by t-SNE differs from SNE in using a student t distribution rather than a Gaussian to compute the similarity in the low dimensional space, which makes it easier to converge during optimization. In order to quantify the difference between the target and source domains analytically, we used KL divergence.

We compared the classification accuracy among multiple machine learning models with same datasets. The models included SVM, MLP, VAE, VAE and MMD combination (VAE+MMD), VAE and ComBat harmonization combination (VAE+ComBat), VAE, ComBat and MMD (VAE+ComBat+MMD), and the final two approaches included transfer learning data (VAE+ComBat+TL, VAE+ComBat+MMD+TL). Accuracy and F1-score were used to measure the performance of the model. Accuracy represents how close the prediction comes to the true values. It is determined by the number of correct predictions divided by the total number of predictions. To compare the target domain training accuracies between approaches, we saved the series of accuracies from all training batches in each VAE model and then applied two sample t-test on two series.

Because of the imbalance in the size of the classes in the test dataset, we used F1-score to combine both precision and recall of each class into the calculation:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

In the multiple class case, the F1 calculates metrics by counting all true positives, false negatives and false positives [273].The range of the F1-score is from 0 to 1, where a value of one indicates perfect precision and recall, and the value of zero indicates that either the precision or the recall is zero. In the context of multiple classes, we calculated the F1-score globally by counting the total true positives, false negatives and false positives.

4.2.10. Feature identification

The first encoding layer is the most interpretable because the weights between each node of the encoding layer to the next hidden layer are considered as learned features [205] [95]. Based on the feature interpretation methods from previous DNN approaches [97] [274], we analyzed the weights from the encoding layer to the next hidden layer to explain the importance of features in the classification.

We applied permutation testing to identify statistically significant features among those identified above. Once the model was trained using training data, the weights assumed values accordingly during the training process. At the end of training, each weight had a mean value calculated over all iterations of training. This mean weight represented the “importance” of the corresponding feature in the input weight vector of size 1×19900. During permutation, the order of the input vector was randomly shuffled and the training process was repeated after each shuffle. The mean weight obtained during each permutation corresponded to the importance of different features in different permutations. The distribution of mean weights obtained across permutations (1000 of them) represented a null distribution of the hypothesis that all features were significantly important. The p-value of node A was calculated by the number of mean weight values greater

than the true value, divided by the total number of mean weight values. The *p*-value was corrected for multiple comparisons using the false discovery rate (FDR) method at 5%. The equation for calculating the p-value was as follows:

$$P_A = \frac{Num_{greatTrue}}{Num_{permutation}} \times 100\%$$

Here, the $Num_{greatTrue}$ refers to the number of permutations where the mean value of weights was greater than the true value and $Num_{permutation}$ refers to the total number of permutation tests (=1000). The permutation testing procedure was identical for all of the models reported.

4.3. Results

4.3.1. Domain adaptation

Figure 19 shows t-SNE visualizations of the latent feature space in both the source (top panel) and target domains (bottom panel), prior to (left panel) and after (right panel) training. Prior to training, there is little separation between the diagnostic groups in both the source and target domains. However, after training, a clear separation of the diagnostic groups in latent space emerges in the source domain. This is transferred to the target domain as well as we see visible separation between diagnostic groups (with some exceptions, especially between autism and Asperger's). Even with high dimensional input data, the VAE-MMD model can achieve robustness by reducing the distance between the data points from same class but different domains in latent space.

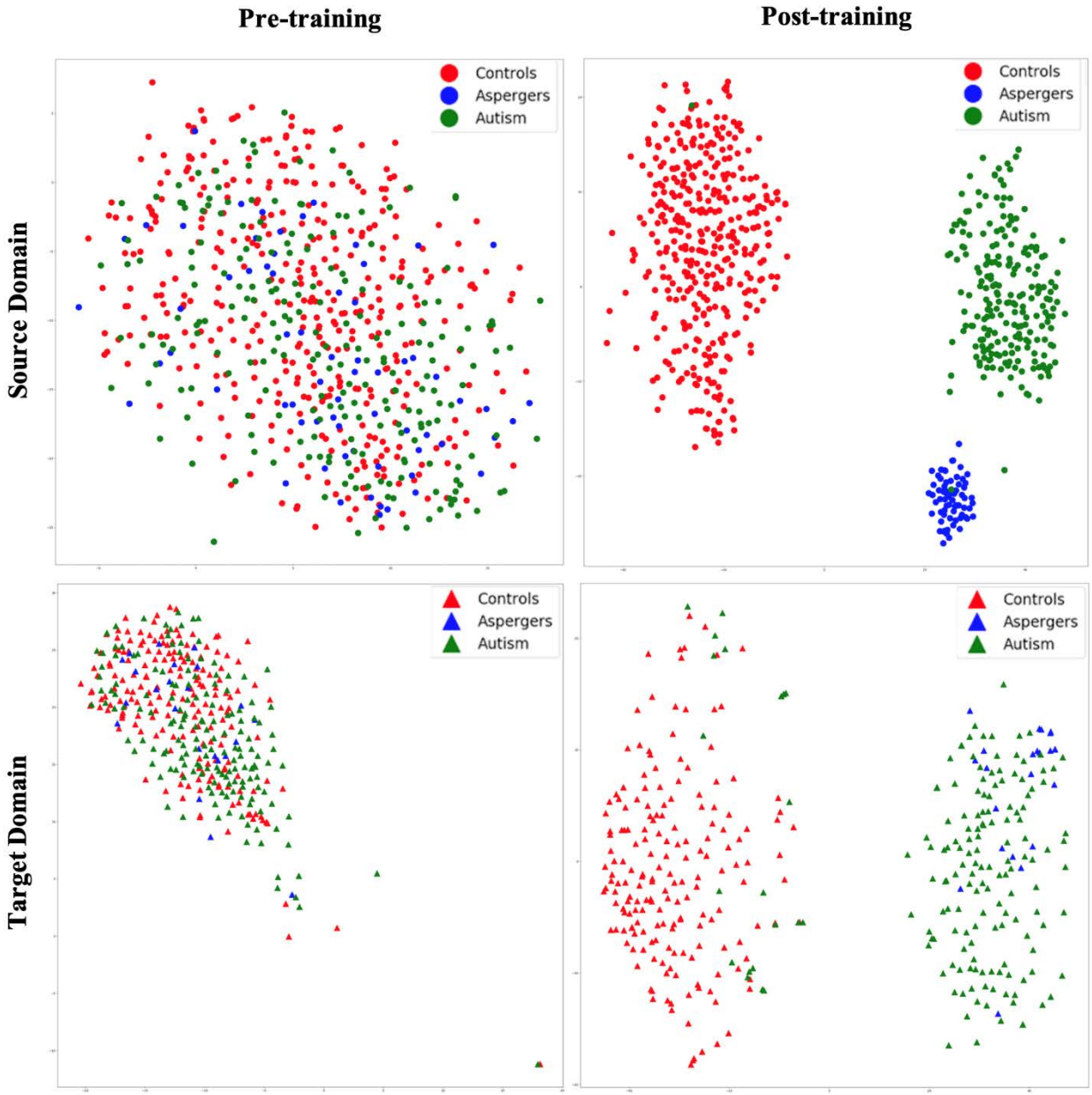

Figure 19: *t-SNE visualization of latent feature spaces for VAE-MMD domain adaptation model. High-dimensional data is reduced to two dimensions for visualization through t-SNE. Left panel: clustering before training; right panel: clustering after training; top panel: source domain; bottom panel: target domain. Red color corresponds to controls whereas blue and green colors correspond to Asperger's and autism patients, respectively. Circle marks correspond to the source domain and triangle marks correspond to the target domain.*

Healthy control subjects from HBN and AOMIC datasets were input as additional source domain data. Since learning about healthy control subjects in one domain (HBN and AOMIC) is

“transferred” to another domain (ABIDE), we refer to this specific case of domain adaptation here as “transfer learning”. The t-SNE embedding (**Figure 20**) shows the latent feature distributions for VAE-MMD domain adaptation model, with transfer learning from additional healthy control data in the source domain drawn from HBN and AOMIC datasets. As with the earlier case, there was very little separation between groups prior to training, in part because the non-neural inter-site differences drown out the inter-group neural differences. After training, we can observe that separation between groups is near perfect in the source domain and visible in the target domain (with some missed assignments to the wrong cluster). Comparing Fig 19 and 20, it is noteworthy that including additional healthy control data in the source domain from HBN and AOMIC datasets expanded the reach of the healthy control cluster in both the source and target domains. This implies that the model was able to capture larger variance in the healthy population, and hence became more generalizable in the target domain as evidenced by improved target domain accuracies presented in the next section. As elaborated in the discussion, it is our hope that with more publicly available healthy control data input into our model in the future, generalizability of the model can be further improved, leading to more realistic separation boundaries between groups, and hence better performance on unseen test data in the target domain.

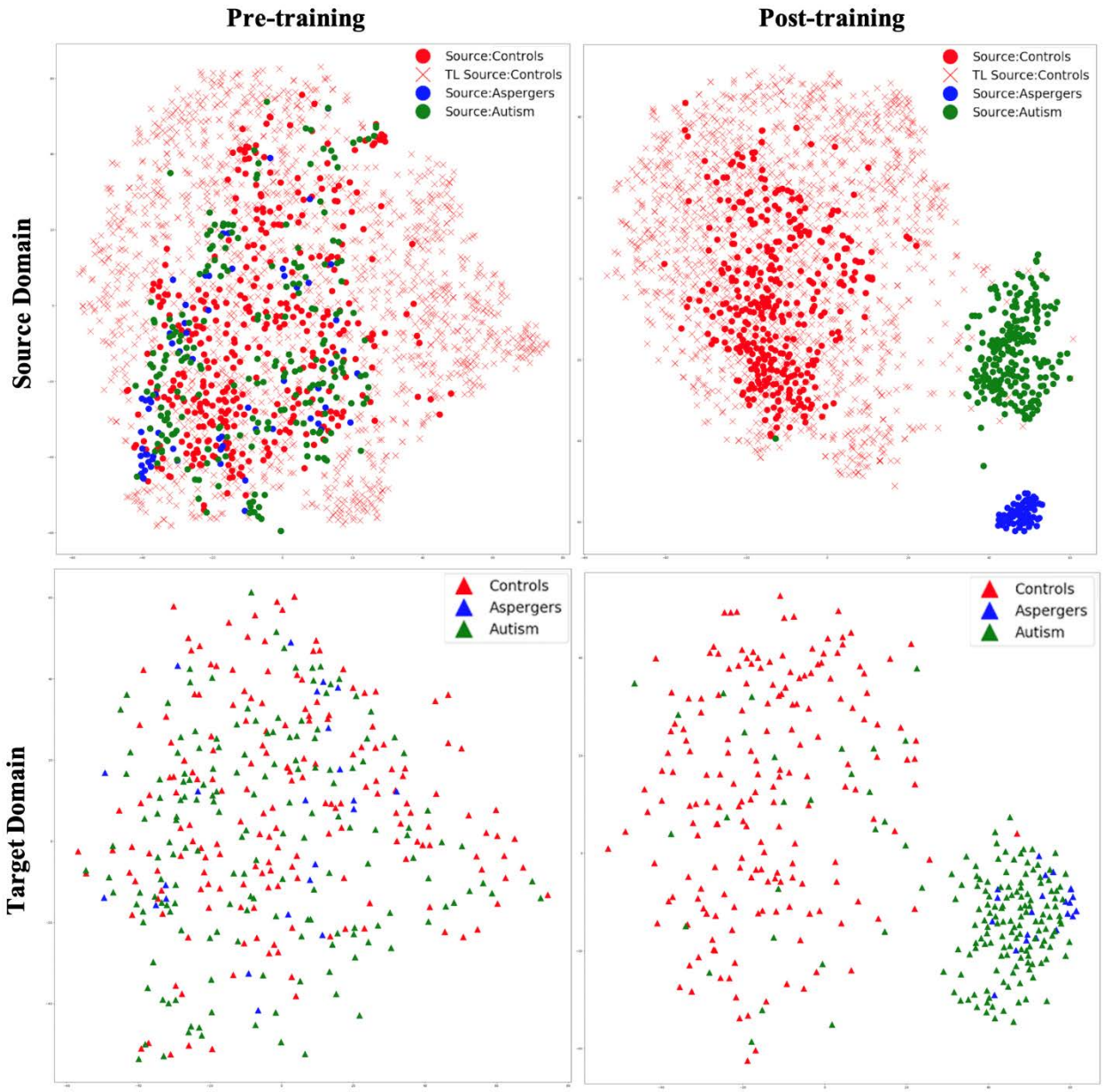

Figure 20: *t-SNE visualization of latent feature spaces for VAE-MMD domain adaptation model, with transfer learning from additional healthy control data in the source domain drawn from HBN and AOMIC datasets. Left panel: clustering before training; right panel: clustering after training; top panel: source domain; bottom panel: target domain. Red color corresponds to controls whereas blue and green colors correspond to Asperger's and autism patients, respectively. Circle marks corresponding to ABIDE I subjects in the original source domain; cross marks correspond to additional HBN and AOMIC healthy control subjects in the source domain and triangle marks corresponding to the target domain.*

4.3.2. Classification accuracy

Table 2 shows the accuracy and F1-score from VAE-MMD model (i.e., domain adaptation) when combined with other strategies such as transfer learning (TL) from HBN and AOMIC datasets, as well as statistical harmonization (ComBat). Also, the results from baseline methods SVM and MLP are included. The baseline methods did not perform well because of the domain shift between source domain data and target domain data. All other methods containing VAE overall performed better than baseline methods. All models have almost 100% training accuracy in the source domain, which is not surprising given that training accuracy tends to saturate. The hypothesis testing results show that VAE+MMD outperformed VAE and VAE+ComBat method in target training accuracy ($p<10^{-5}$). The VAE+MMD+ComBat achieved higher accuracy than three individual methods ($p<10^{-6}$). VAE+MMD+ComBat, VAE+MMD+TL and VAE+MMD+ComBat+TL approaches have no significant accuracy difference ($p>0.05$). The testing accuracy of the source domain is poor given the inability to generalize based just on the source data. This agrees with prior results on the application of standard machine learning methods to neuroimaging data [115]. MMD-based domain adaptation enhances the accuracy during target domain training. For the three classes in the target domain dataset (controls, autism and Asperger's), MMD domain adaptation can increase accuracy by 4% to 10% without using target domain labels. Combining MMD domain adaptation with other methods such as transfer learning and/or statistical harmonization (ComBat) can further boost training and testing accuracy as well as the F1-scores in the target domain (**Table 3**, **Figure 21**).

For three-way classification using test data in the target domain, our model achieved 74.6% accuracy with domain adaptation and ComBat. Given the smaller number of samples from Asperger's and its similarities with autism, three-way classification in ABIDE is a hard problem.

Although cross-validation accuracies above chance (which is 33%) have been reported before, accuracy in independent test datasets rarely exceeded 70% [275] [213] [115]. If we included AOMIC and HBN datasets into the source domain, accuracy further increased to 75.4% due to transfer learning, demonstrating that there is scope within the domain adaptation framework to improve the accuracy further by including more data. Considering differences in data distribution between AOMIC and HBN datasets (Table 1), we investigated how much improvement in performance were caused by HBN and AOMIC separately. Table 3 shows the accuracy results from separately including HBN and AOMIC datasets into the VAE+ComBat+MMD+TL framework.

Table 3: *Classification results obtained by combining domain adaptation (VAE-MMD) with other strategies such as transfer learning (TL) and statistical harmonization (ComBat). For each of the training and testing datasets (test sample size equal to 126), we compared the classification accuracies from source and target domains. In addition, we used SVM, MLP and VAE trained on source domain data and tested on target domain data. The last column shows the F1-score of each approache. Domain adaptation was not applied with SVM and MLP, so there was no source training, source test and target training classification accuracies.*

Classification Accuracy	Source Training	Source Test	Target Training	Target Test (F1-score)
SVM	~	~	~	49.32% (0.32)
MLP	~	~	~	62.66% (0.30)
VAE	99.97%	64.56%	52.67%	65.08% (0.27)
VAE+COMBAT	100%	60.76%	51.56%	70.63% (0.29)
VAE+MMD	99.67%	50.94%	63.29%	69.05% (0.35)
VAE+MMD+COMBAT	100%	52.53%	83.11%	74.6% (0.47)
VAE+MMD+TL	100%	60.76%	77.61%	73.81% (0.38)
VAE+MMD+COMBAT +TL	98.79%	52.53%	82.11%	75.4% (0.44)

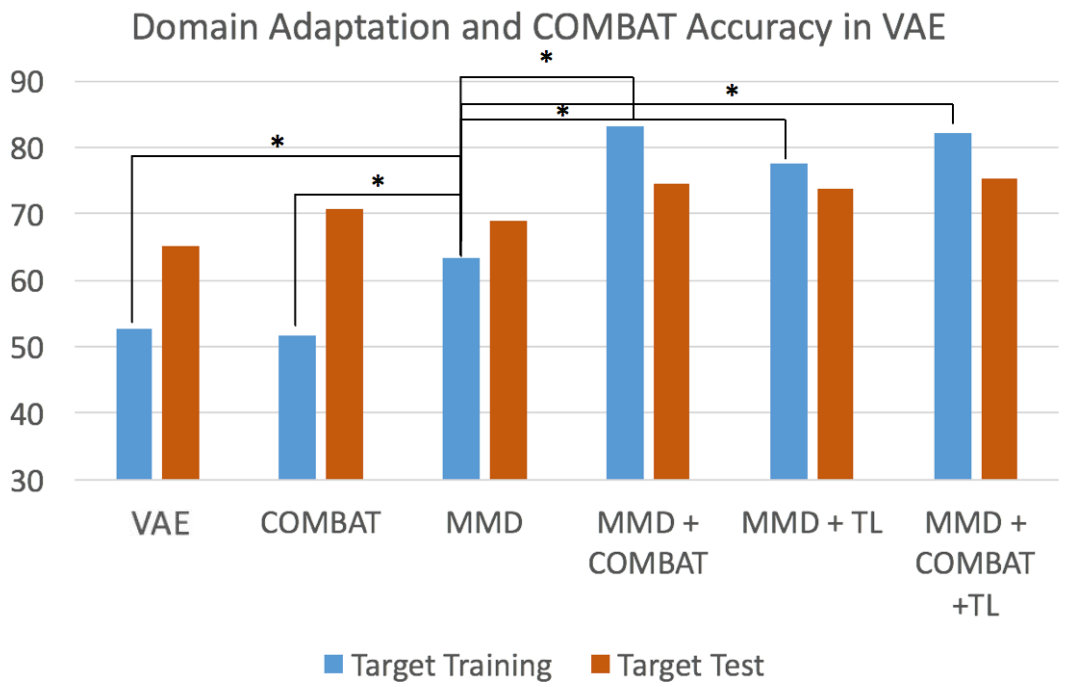

Figure 21: *The classification accuracy using different approaches combined with domain adaptation and ComBat harmonization. Blue bars refer to the training accuracy in the target domain and red bars refer to the testing accuracy in the target domain. The error bars with star reveal the significant differnce between the accuracies($p<0.05$). Some obvious comparisons are not shown in this figure to keep the figure concise.*

Table 4. Classification results obtained by including AOMIC and HBN data in the model separately. While age and gender composition of HBN were comparable to ABIDE, the AOMIC cohort was older with more proportion of females (please refer to Table 1).

Classification Accuracy	Source Training	Source Test	Target Training	Target Test (F1-score)
VAE+MMD+COMBAT + AOMIC	99.18%	56.96%	80.11%	74.6% (0.47)
VAE+MMD+COMBAT + HBN	97.30%	59.49%	82.06%	75.4% (0.44)

4.3.3. Feature identification

Figure 22 shows the features important for classification using the VAE+MMD+TL+ComBat model. These paths also happen to be significantly weaker ($p<0.05$, FDR corrected) [276] in ASD

and Asperger's as compared to healthy controls. Except the local connection of supramarginal gyrus to postcentral gyrus in the parietal lobe, most of the paths were cross-network and cross-lobe connections including middle frontal gyrus to inferior temporal gyrus, and BA6 to middle temporal gyrus in the fronto-temporal network, orbito-frontal gyrus to rolandic operculum in the fronto-insular network, and right precentral to right temporal pole in the temporo-parietal network. Most of the affected regions in the frontal lobe were left lateralized.

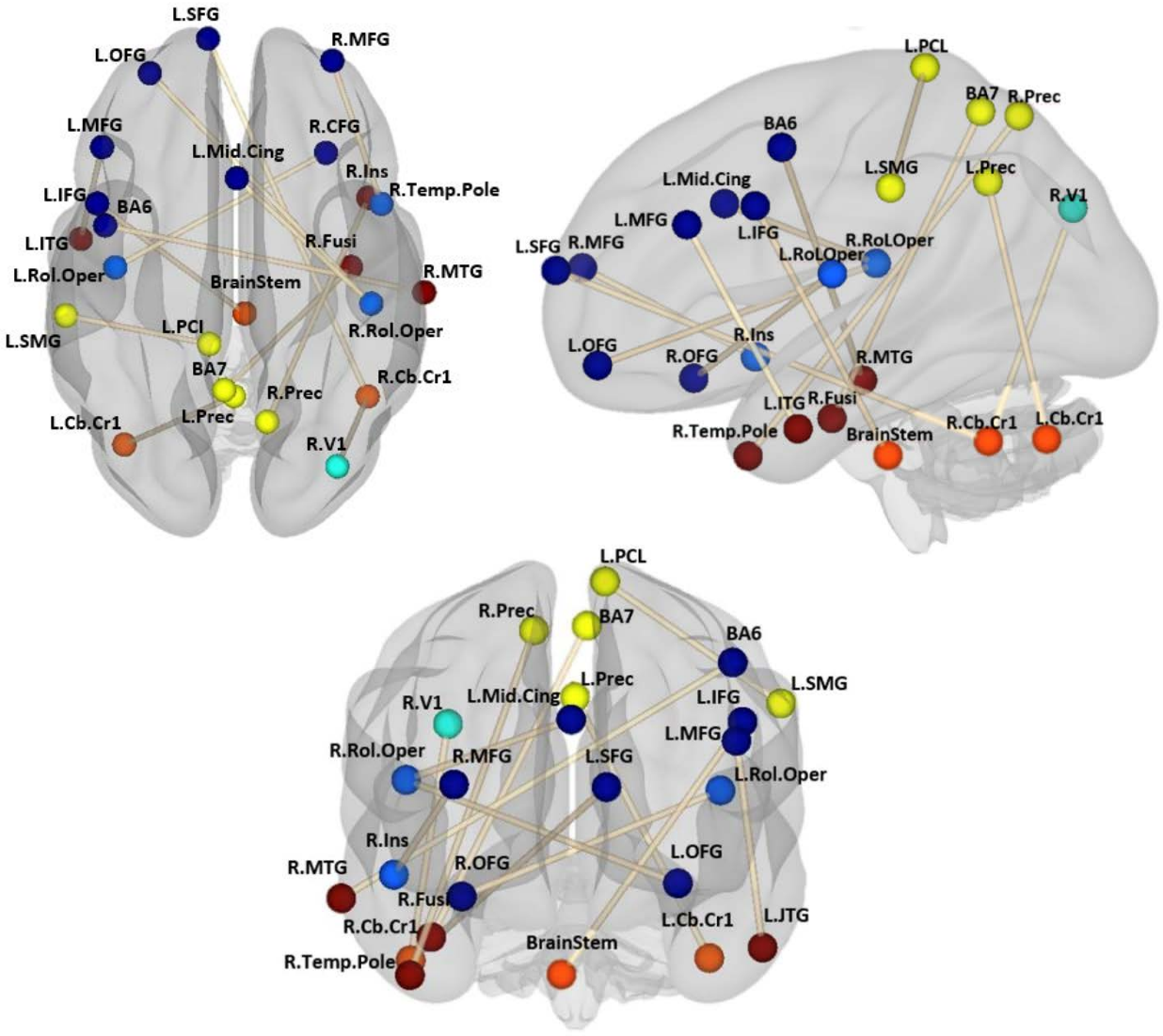

Figure 22: *FC features found to be important for classification using our VAE+MMD+TL+ComBat model with highest target testing accuracy. Figure shows coronal, sagittal and axial views of connections with colormap. The colors represent different lobes: dark blue: frontal lobe; light blue: insular lobe; cyan: occipital lobe; yellow: parietal lobe; orange: subcortical; red: temporal lobe.*

4.4. Discussion

Large public neuroimaging databases have kindled the possibility of using deep learning models for diagnostic classification with potential applications in AI-assisted clinical decision systems. However such large public databases have been assembled post-hoc, and hence contain different sources of non-neural variability such as different sites using different scanners and protocols, which degrade performance of deep learning as well as traditional machine learning models [213] [115]. To address this, we proposed and implemented a domain adaptation framework employing a VAE-MMD deep learning model using ABIDE I as the source domain and ABIDE II as the target domain. We demonstrated improved classification performance in the target domain by utilizing the knowledge learned from the source domain by making the distributions of data in source and target domains as similar as possible [216] [217]. The ComBat statistical harmonization [248] [249] when used in combination with domain adaptation improved the performance of the classifier. We also showed that additional transfer learning from HBN and AOMIC datasets further improved the classification accuracy.

Even with high dimensional input features, the VAE-MMD model was able to project data points from different domains from the same class into a closed latent space. Our results demonstrate that deep learning based transfer learning and domain adaptation as well as statistical ComBat harmonization approaches can improve target domain classification when used independently. When used in combination, the accuracy was better than when they were used independently, indicating that both statistical and deep learning methods bring something unique

to the table. Specifically, Figure 21 and Table 3 showed that learning from labeled training data in the source domain improved dramatically with domain adaptation and ComBat harmonization, with the same trends seen in the target domain with unlabeled data, but to a lesser extent [277]. Compared to these two methods (ComBat harmonization and domain adaptation approach), ComBat requires minimal hardware and time to complete the harmonization, while the deep learning model has more hyper-parameters for fine-tuning and is harder to train. It remains to be seen whether the improvement in performance expected by including larger datasets in the deep learning framework will justify the additional computational complexity in comparison with statistical ComBat harmonization.

In our study, a three-class classification approach was used (Autism, Asperger's syndrome and Controls). The Asperger's population is more similar to autism than healthy controls [278], but is still distinctly separate from typical autism across behavioral, cognitive and neural domains [279]. Although several studies prefer to perform two-way classification between controls and ASD, some studies have performed three-way classification as well [280]. While three-way classification performance reported in the relatively larger ABIDE dataset has been modest, relatively good three-way classification accuracy of deep learning versus traditional machine learning models in the case of ASD (over 70% accuracy) has been reported in smaller datasets (N =114) [281]. We found that the VAE+MMD+ComBat+TL approach outperformed SVM and MLP methods by enhancing classification of the Asperger's class from less than 10% to about 60%. One of the three-way ASD classification studies [280] also applied a domain adaptation approach and used functional connectivity as input features, but they reported less than 60% accuracy in ASD classification. Thus, compared to other three-way ASD classification

studies, our approach obtained a high test accuracy of over 75% (Table 3). The Table 5 below shows that the domain adaptation approach shows better two-way classification performance compared to other machine learning methods. Compared to the three-way results in the main paper, we found the accuracy improvement from SVM (+15%) and MLP (+8%) is more than the accuracy improvement from the domain adaptation approach (+1%). In a way, this indicates that the domain adaptation based deep learning approach is not very sensitive to the number of classes. It also shows that the advantages provided by this approach will become more substantial in multi-class classificaion problems.

Table 5. *Two-way classification results obtained by combining domain adaptation (VAE-MMD) with other strategies such as transfer learning (TL) and statistical harmonization (ComBat) compared to SVM and MLP . We compared the binary training and testing classification accuracies from source and target domains. The last column shows the F1-score of each approache in target test data.*

Classification Accuracy	Source Training	Source Test	Target Training	Target Test (F1-score)
SVM	~	~	~	64.28% (0.27)
MLP	~	~	~	70.63% (0.29)
VAE+MMD+COMBAT +TL	98.50%	56.96%	86.33%	76.19% (0.47)

As discussed in the introduction, our semi-supervised domain adaptation approach combined the advantages of UDA and SDA. In prior literature, one application of UDA focused on sub-cortical and lesion segmentation using a CNN model to extract domain-transferable features, accounting for differences in MRI scanners and image acquisition parameters [106]. Another UDA approach has been used for brain lesion segmentation to achieve high segmentation accuracy compared to SDA approaches [282]. In this study, without the annotated labels in the target domain, our approach is the first to utilize such a UDA framework in a psychiatric disorder classification task [232]. In the absence of such inter-site harmonization efforts, the inter-class differences are

drowned in sizeable inter-site variability. During the domain adaptation process, t-SNE helped us visualize how domain adaptation improved classification between groups in the latent space [283]. Previous neuroimaging studies have used t-SNE as a visualization method to validate the domain adaptation result [284] [285] [42], but this study was the first to use it in psychiatric disorder prediction. Compared to other SDA studies, our approach provided higher accuracy in ASD classification. Specifically, the most comparable study is by Shi et al [286], wherein they trained the three-way decision domain adaptation classifier with the MMD model, which was applied to FC from the ABIDE dataset, and obtained around 71% accuracy. They used propagated pseudo labels to target domain data trained by an SVM classifier, which does not benefit from a deep learning classifier to handle high-dimensional data as in our model. Different from Shi et al [286], another study [287] treated one individual site as a target domain and all other sites as source domains. Then, a common low-rank latent representation was constructed across the source and target domains, obtaining 60% to 70% accuracy. Thus, our approach (at 75.4% accuracy) yields superior performance over these state-of-the-art domain adaptation methods applied to ASD prediction.

Since domain adaptation improved target domain test accuracy, it raises the possibility that additional data in the source domain may further improve classification performance. However, we reasoned that large additional data from disorders (in this case ASD) takes relatively more time to become available in the public domain. Therefore, it makes sense to explore whether further improvement in target test accuracy could be achieved by augmenting the source domain with additional healthy control data. There is already considerable amount of healthy control data in the public domain. Therefore, if the classifier can better generalize the underlying brain patterns corresponding to healthy controls, then it must naturally result in better discrimination of healthy

controls from ASD in the target domain [265]. With this logic in mind, we augmented the source domain with additional healthy control data from HBN and AOMIC datasets. From Table 2, it is clear that AOMIC has a higher proportion of females and is older in mean age compared to the other three datasets. Despite this, we chose AOMIC with the intention of improving the generalizability of the classifier (by exposing the classifier to different age/gender mixes). The results from separate datasets are shown in Table 3; HBN provided slightly better performance than AOMIC, likely because it has similar age and gender to ABIDE. Combined both AOMIC and HBN datasets, transfer learning from these datasets to discrimination in the target domain did predictably improve performance. An outstanding question is whether further improvements in performance can be achieved by augmenting the source domain with additional healthy control data from open datasets such as the UK biobank [288], brain genomics superstruct project [289], Nathan Kline Rockland sample [290], 1000 connectomes [291] and Philadelphia Neurodevelopmental Cohort [292], etc. While investigating this question is beyond the scope of the current report, it is a tantalizing possibility that needs to be probed. If additional data does improve performance further, it opens up the possibility of building classifiers that are truly generalizable to the population at large. This is an essential step in making machine learning models based on neuroimaging data relevant for AI-based diagnostic support in the clinic, rather than being a purely academic tool (which it is currently) for understanding discriminative features of brain function in mental disorders.

The connectivities identified via feature importance analysis displayed hypo-connectivity in ASD, and this is highly consistent with previous ASD research. Previous deep learning approaches have identified ROIs such as right superior frontal gyrus, right middle temporal gyrus [293], rolandic operculum, and insula [251] as being discriminative for ASD. These same regions have

also been implicated in our study. Most identified functional connections were associated with regions in fronto-insular, fronto-temporal, temporo-parietal and fronto-parietal axes. The marked functional hypo-connectivity patterns of the fronto-temporal pathway may underlie language impairment in children with ASD [294]. Previous work involving working memory for faces found lower activation in a right temporal area in ASD relative to controls, as well as a somewhat different location of the activation in the fusiform area [295] [296] [297]. Also, findings from previous statistical analysis studies strongly implicate atypical between-network FC of the frontal-parietal pathway as possible brain markers of ASD [298] [299]. Even three decades ago, researchers had demonstrated that individuals with ASD showed reduced connectivity between the insula and fronto-parietal regions [300]. Decreased connectivity in ASD subjects in temporo-parietal networks during object recognition has also been observed [298]. Our result also shows several decreased anterior-posterior functional connectivity paths, which validates results reported previously [301].

Chapter 5

Conclusion

5.1 Conclusions

The results from the first study show that the TL method outperforms ASD classification in test data as compared to other state-of-the-art methods, especially with small training sample size. The convolutional VAE model is robust against data divergence across different sites and between different data sources. We have demonstrated the applicability of transfer learning within a deep learning framework for utilizing larger samples of available healthy control data to improve generalizability and accuracy of diagnostic classification in ASD, as well as reduce the harmful effects of inter-site variability on classification. In addition, we identified FC patterns in the brain network, which corroborated with previously published FC impairments in ASD. In summary, for both site matched and mismatched splits, TL using a VAE-CNN classifier outperforms a traditional deep learning CNN classifier without TL. Notably, the accuracies we have obtained on the independent test data are higher than those obtained from cross-validation in ABIDE [132] [163] (even though cross-validation over-estimates performance [115]), as well as on independent test data using traditional machine learning [115] and other TL approaches [132]. TL seems to alleviate deterioration of performance due to site-mismatched split, which may help in multi-site studies. Widespread hypo-connectivity in inter-hemispheric and inter-

lobar connections with sporadic (local) hyper-connectivity in the right hemisphere (especially frontal lobe) in ASD are in line with previous results in the literature [211] [302] [303].

In the second study, our results demonstrate that domain adaptation and transfer learning, when used individually or in combination with statistical harmonization techniques such as ComBat, outperforms ASD classification in test data as compared to baseline methods without using any harmonization of multi-site data. The domain adaptation VAE-MMD model is robust against sources of data distribution divergence such as inter-site differences in data acquisition parameters and scanner models. By demonstrating that augmenting the source domain with additional data leads to improved target domain accuracy due to transfer learning, our work opens the possibility of further improvement of the model by utilizing the ever-increasing amount of healthy control neuroimaging data in the public domain.

5.2 Limitations and Future Work

The limitations of the first study including while generating the synthetic data, we applied a basic VAE model to generate data using limited training data. Generative Adversarial Network (GAN) is a more popular data generation model in computer vision, drug engineering [146], as well as in imaging [304]. The applicability of GANs in this context is worth exploring in the future. Next, we used over 1000 subjects from the HCP dataset for transfer learning. Some limitations of the second study must be kept in mind while interpreting the results. First of all, we have not explored the best interpretability algorithms for the type of machine learning model we have used. This is a topic for future research. Especially for the transferred features across different data distributions, there are several studies using reference models to obtain successful learning representations of domain invariant features [240] [216]. Second, we have not investigated when the benefit from domain adaptation and transfer learning saturates. This requires additional

samples from either/both groups, which is increasingly becoming feasible given the open availability of data these days. Third, how dependent is the performance of the framework on the inherent heterogeneity of the (i) sample, (ii) disorder, (iii) data acquisition and preprocessing strategies needs to be investigated further.

In the future work, first of all, there are many other large public fMRI datasets that are currently available such as the UK Biobank [80] (N>40,000), the adolescent brain cognitive development (ABCD) [81] (N=1,1975), the healthy brain network (HBN) [136](N=2505), etc. Those databases have data acquired from subjects with varied ages, gender, race, and other demographics that can enhance the generalizability of the training model. It is yet unclear whether including these large datasets in the training model will boost performance, or whether the performance saturates at some point. This is a topic ripe for future research. Second, how the proposed approach interacts with similar approaches is unknown, i.e. when we combine different approaches used to counter inter-site variability in the data and enhance generalizability, are the performance enhancements additive or do cascading methods provide diminishing returns? For example, ComBat harmonization [245] is a statistical technique used to normalize the variability of data distributions arising from different sources of data (such as acquisition site) so that they can be pooled together to remove the non-neural variability across scanners and sites [248]. Third, to enhance the performance of the multi-class approach, the class imbalance issue needs to be further addressed. As found in an earlier ASD study [305], Synthetic Minority Over-sampling Technique (SMOTE) [306] is a series of effective data argumentation methods for oversampling the data in the minority class of imbalanced datasets such as Asperger's syndrome. As opposed to this synthetic data method, reverse domain adaptation is also a promising domain adaptation

solution to translate the real data to synthetic data [307]. These sophisticated methods may better handle the class imbalance that is often prevalent in three-way classification of Autism, Asperger's and controls.

Reference

[1] S. Ogawa, T. Lee, A. Kay and D. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proc Natl Acad Sci USA,* p. 87:9868–72, 1990.

[2] E. M. Hillman, "Coupling Mechanism and Significance of the BOLD Signal: A Status Report," *Annual Review of Neuroscience,* vol. 37, no. 1, pp. 161-181, July 2014.

[3] R. Buxton, "Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques," *Energy,* vol. 24, no. 2, p. 523, 2002.

[4] S.-G. Kim and K. Ugurbil, "Functional magnetic resonance imaging of the human brain," *Neurosci. Methods,* vol. 74, no. 2, pp. 229-243, 1997.

[5] N. Logothetis, J. Pauls, M. Augath, T. Trinath and Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature,* vol. 412, pp. 150-7, 12 July 2001.

[6] M. Blaschko, J. Shelton and A. Bartels, "Augmenting Feature-driven fMRI Analyses: Semi-supervised learning and resting state activity.," in *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference,* 2009.

[7] B. B. Biswal, J. V. Kylen and J. S. Hyde, "Simultaneous assessment of flow and bold signals in restingstate functional connectivity maps," *NMR Biomed,* vol. 10, no. 4-5, pp. 165-170, 1997.

[8] M. Bianciardi, M. Fukunaga, P. van Gelderen, S. G. Horovitz, J. A. d. Zwart and J. H. Duyn, "Modulation of spontaneous fMRI activity in human visual cortex by behavioral state," *NeuroImage,* vol. 45(1), pp. 160-168, 2009.

[9] A. Akhavan, M. A. Sharifi and M. M. Pedram, "Combination of rs-fMRI and sMRI Data to Discriminate Autism Spectrum Disorders in Young Children Using Deep Belief Network," *J Digit Imaging,* vol. 31, p. 895–903, 2018.

[10] H. Karbasforoushan and N. Woodward, "Resting-State Networks In Schizophrenia," *Current topics in medicinal chemistry,* vol. 12, pp. 2404-2414, 2012.

[11] S. H. Hojjati, "Identification of the Early Stage of Alzheimer's Disease Using Structural MRI and Resting-State fMRI ," *Frontiers in neurology,* vol. 10, p. 904, 30 8 2019.

[12] H. A. Jaber, H. K. Aljobouri, I. Çankaya, K. O. M. and O. Algin, "Preparing fMRI Data for Postprocessing: Conversion Modalities, Preprocessing Pipeline, and Parametric and Nonparametric Approaches," *IEEE Access,* vol. 7, pp. 122864-122877, 2019.

[13] C. Strother Stephen, "Evaluating fMRI Pre- processing Pipelines: Review of Preprocessing Steps for BOLD fMRI," *IEEE Eng Med Biol Mag,* vol. 25, no. 2, pp. 27-41, 2006.

[14] D. Power Jonathan, A. Barnes Kelly, Z. Snyder Abraham, L. Schlaggar Bradley and E. Petersen Steven, "Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion," *NeuroImage,* vol. 59, pp. 2142-2154, 2012.

[15] W. Wu Changwei, C.-L. Chen, P.-Y. Liu, Y.-. P. Chao, B. Biswal Bharat and C.-P. Lin, "Empirical Evaluations of Slice-Timing, Smoothing, and Normalization Effects in Seed-Based, Resting-State Functional Magnetic Resonance Imaging Analyses," *Brain Connectivity,* vol. 1, pp. 401-410.

[16] Maknojia and Sanam, *Resting State fMRI: Going Through the Motions* , vol. 13, p. 825, 13 8 2019.

[17] W. Huijbers, K. R. A. Van Dijk, M. M. Boenniger, R. Stirnberg and M. M. B. Breteler, "Less head motion during MRI under task than resting-state conditions," *Neuroimage,* vol. 147, pp. 111-120, 2017.

[18] A. V. Oppenheim, "Applications of Digital Signal Processing," *Prentice Hall,* 1978.

[19] T. Alakörkkö, H. Saarimäki, E. Glerean, J. Saramäki and O. Korhonen, "Effects of spatial smoothing on functional brain networks," *The European Journal of Neuroscience,* vol. 46(9), p. 2471–2480, 2017.

[20] E. Molloy, M. Meyerand and R. Birn, "The influence of spatial resolution and smoothing on the detectability of resting-state and task fMRI," *NeuroImage,* vol. 86, pp. 221-230, 2014.

[21] X.-N. Zuo, T. Xu, L. Jiang, Z. Yang, X.-Y. Cao, Y. He, Y.-F. Zang and F. Castellanos, "Toward reliable characterization of func- tional homogeneity in the human brain: preprocessing, scan duration, imaging resolution and computational space," *NeuroImage,* vol. 65, pp. 374-386, 2013.

[22] K. Aquino, P. Robinson, M. Schira and M. Breakspear, "Deconvolution of neural dynamics from fMRI data using a spatiotemporal hemodynamic response function," *NeuroImage,* vol. 94, 10 3 2014.

[23] J. T. Webb, M. A. Ferguson, J. A. Nielsen and J. S. Anderson, "BOLD Granger causality reflects vascular anatomy," *PLoS One,* vol. 8, no. 12, p. e84279, 2013.

[24] O. David, I. Guillemain, S. Saillet, S. Reyt, C. Deransart, C. Segebarth and A. Depaulis, "Identifying Neural Drivers with Functional MRI: An Electrophysiological Validation," *PLoS Biol,* vol. 6, no. 12, p. e315, 2008.

[25] C. Chang, M. Thomason and G. Glover, "Mapping and correction of vascular hemodynamic latency in the BOLD signal," *Neuroimage,* vol. 43, p. 90–102, 2008.

[26] K. Friston, N. Trujillo-Barreto and J. Daunizeau, "Dem: a variational treatment of dynamic systems," *Neuroimage,* vol. 41, p. 849–885, 2008.

[27] G.-R. G. Wu, W. Liao, S. Stramaglia, J.-R. J. Ding, H. Chen and D. …. Marinazzo, "A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data," *Med. image,* vol. 17, no. 3, p. 365–74, 2013.

[28] M. Raichle, A. MacLeod, A. Snyder, W. Powers, D. Gusnard and G. Shulman, "A default mode of brain function," *Proc Natl Acad Sci USA,* vol. 98, p. 676–682, 2001.

[29] G. Shulman, J. Fiez, M. Corbetta, R. Buckner, F. Miezin, M. Raichle and S. Petersen, "Common blood flow changes across visual tasks. I. Increases in subcortical structures and cerebellum, but not in non-visual cortex," *J Cogn Neurosci,* vol. 9, p. 648–663, 1997.

[30] B. Rogers, V. Morgan, A. Newton and J. Gore, "Assessing Functional Connectivity in the Human Brain by FMRI," *Magnetic resonance imaging,* vol. 25, pp. 1347-57, 1 1 2008.

[31] M. Stevens, G. Pearlson and V. Calhoun, "Changes in the Interaction of Resting-State Neural Networks From Adolescence to Adulthood," *Human brain mapping,* vol. 30, pp. 2356-66, 08 01 2009.

[32] D. Cordes, V. Haughton, K. Arfanakis, G. Wendt, P. Turski, C. Moritz, M. Quigley and M. Meyerand, "Mapping functionally related regions of brain with functional connectivity MRI (fcMRI)," *AJNR.American journal of neuroradiology,* vol. 21, pp. 1636-44, 11 2000.

[33] N. Woodward and C. Cascio, "Resting-State Functional Connectivity in Psychiatric Disorders," *JAMA psychiatry,* vol. 72, 6 10 2015.

[34] M. A. Just, T. A. Keller, V. L. Malave, R. K. Kana and S. Varma, "Autism as a neural systems disorder: A theory of frontal- posterior underconnectivity," *Neuroscience & Biobehavioral Reviews,* vol. 36, p. 1292–1313, 2012.

[35] M. K. Belmonte, G. Allen, A. Beckel-Mitchener, L. M. Boulanger, R. A. Carper and S. J. Webb, "Autism and abnormal development of brain connectivity," *Journal of Neuroscience,* vol. 24, no. 42, p. 9228–9231, 2004.

[36] A. S. Nunes, N. Peatfield, ,. Vakorin and S. M. Doesburg, "Idiosyncratic organization of cortical networks in autism spectrum disorder," *NeuroImage,* 2018.

[37] A. Easson, Z. Fatima and A. McIntosh, "Functional connectivity-based subtypes of individuals with and without autism spectrum disorder," *Network Neuroscience,* vol. 3, pp. 1-43, 2018.

[38] J. Baker, A. Holmes, G. Masters, B. T. Yeo, F. Krienen, R. Buckner and D. Ongür, "Disruption of Cortical Association Networks in Schizophrenia and Psychotic Bipolar Disorder,," *JAMA psychiatry (Chicago, Ill.),* vol. 71.

[39] B. Monroe and B. Todd, "The Impact of Alzheimer's Disease on the Resting State Functional Connectivity of Brain Regions Modulating Pain: A Cross Sectional Study," *Journal of Alzheimer's disease : JAD,* vol. 57, no. 1, pp. 71-83, 2017.

[40] M. Arbabshirani, S. Plis, J. Sui and V. Calhoun, "Single subject prediction of brain disorders in neuroimaging: promises and pitfalls," *Neuroimage,* p. 137–165, 2016.

[41] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature,* vol. 521, p. 436–444, 2015.

[42] S. M. Plis, D. R. Hjelm, R. Salakhutdinov, E. A. Allen, H. J. Bockholt and J. D. Long, "Deep learning for neuroimaging: a validation study," *Front Neurosci,* vol. 8, pp. 1-11, 2014.

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

[44] 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.," *Frontiers in Human Neuroscience,* vol. 7, p. 599, 2013.

[45] J. Donahue, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 4, pp. 677-691, 1 4 2017.

[46] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 4945-4949, 2016.

[47] A. Graves, A.-r. Mohamed and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings.* , 2013.

[48] D. Durstewitz, G. Koppe and A. Meyer-Lindenberg, "Deep neural networks in psychiatry," *Molecular Psychiatry,* vol. 24, p. 1, 1 2 2019.

[49] D. Wen, Z. Wei, Y. Zhou, G. Li, X. Zhang and W. Han, "Deep Learning Methods to Process fMRI Data and Their Application in the Diagnosis of Cognitive Impairment: A Brief Overview and Our Opinion," *Front Neuroinform,* pp. 12-23, 2018.

[50] S. M. G. Vieira, W. H. L. Pinaya and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neuroscience and Biobehavioral Reviews," *Neuroscience & Biobehavioral Reviews,* vol. 74, pp. 58-75, 3 2017.

[51] E. Lin, P. Kuo, Y. Liu, Y. Yu, A. Yang and S. Tsai, "A Deep Learning Approach for Predicting Antidepressant Response in Major Depression Using Clinical and Genetic Biomarkers," *Front Psychiatry,* p. 9:290, 2018.

[52] K. Oh, Y. Chung and K. Kim, "Classification and Visualization of Alzheimer's Disease using Volumetric Convolutional Neural Network and Transfer Learning," *Sci Rep,* vol. 9, p. 18150, 2019.

[53] H. Jang, S. Plis, V. Calhoun and J.-H. Lee, "Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks," *NeuroImage,* vol. 145, 1 4 2016.

[54] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan and C. Zhang, "Learning efficient convolutional networks through network slimming.," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint,* 2014 arXiv:1409.1556.

[56] C. Dansereau, Y. Benhajali and C. Risterucci, "Statistical power and prediction accuracy in multisite resting-state fMRI connectivity," *Neuroimage,* pp. 149:220-232, 2017.

[57] O. Demirci, V. P. Clark, V. A. Magnotta, N. C. Andreasen, J. Lauriello, K. A. Kiehl, G. D. Pearlson and V. D. Calhoun, "A Review of Challenges in the Use of fMRI for Disease

Classification / Characterization and A Projection Pursuit Application from Multi-site fMRI Schizophrenia Study.," *Brain imaging and behavior,* vol. 2(3), p. 147–226, 2008.

[58] A. Di Martino, D. O'Connor and B. Chen, "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. ," *Sci Data,* vol. 4, p. 170010, 2017.

[59] J. A. Turner, E. Damaraju, T. G. van Erp, D. H. Mathalon, J. M. Ford, J. Voyvodic, B. A. Mueller, A. Belger, J. Bustillo, S. McEwen, S. G. Potkin, Fbirn and V. D. Calhoun, "A multi-site resting state fMRI study on the amplitude of low frequency fluctuations in schizophrenia," *Frontiers in neuroscience,* vol. 7, p. 137, 2013.

[60] A. Payan and G. Montana, "Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks.," in *ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods*, 2015.

[61] M. Khosla, K. Jamison, A. Kuceyeski and M. R. Sabuncu, "Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. ," *NeuroImage,* vol. 199, p. 651–662, 2019.

[62] R. Meszlényi, K. Buza and Z. Vidnyánszky, "Resting State fMRI Functional Connectivity-Based Classification Using a Convolutional Neural Network Architecture," *Frontiers in Neuroinformatics* , vol. 11, 2017.

[63] Y. Li, J. Liu, J. Huang, Z. Li and P. Liang, "Learning Brain Connectivity Sub-networks by Group- Constrained Sparse Inverse Covariance Estimation for Alzheimer's Disease," 2018.

[64] X. Chen, j. He, R. Lawrence and J. Carbonell, "Adaptive Multi-task Sparse Learning with an Application to fMRI Study," 2018.

[65] C. Garbin, X. Zhu and O. Marques, "Dropout vs. batch normalization: an empirical study of their impact to deep learning," *Multimedia Tools and Applications*, pp. 1-39.

[66] G. Mesnil, Y. Dauphin, X. Glorot, S. Rifai, Y. Bengio, I. Goodfellow, E. Lavoie, X. Muller, G. Desjardins, D. Warde-Farley, P. Vincent, A. Courville, J. Bergstra, I. Editor, G. Guyon, V. Dror, G. Lemaire, D. Taylor and D. Silver, "Unsupervised and Transfer Learning Challenge: a Deep Learning Approach," vol. 7, pp. 1-15, 2011.

[67] C. R. Phang, F. Noman, H. Hussain, C. M. Ting and H. & Ombao, "A Multi-Domain Connectome Convolutional Neural Network for Identifying Schizophrenia From EEG Connectivity Patterns ," *IEEE journal of biomedical and health informatics,* vol. 24(5), p. 1333–1343, 2020.

[68] A. Di Martino, C. G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart and M. P. Milham, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism.," *Molecular psychiatry*, vol. 19, no. 6, p. 659–667, 2014.

[69] A. Di Martino, D. O'Connor and B. Chen, "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II," *Sci Data* , vol. 4, p. 170010, 2017.

[70] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub and K. Ugurbil, "The WU-Minn Human Connectome Project: an overview," *NeuroImage,* vol. 80, p. 62–79, 2013.

[71] K. Kalcher, W. Huf, R. Boubela, P. Filzmoser, L. Pezawas, B. Biswal, S. Kasper, E. Moser and C. Windischberger, "Fully Exploratory Network Independent Component Analysis of the 1000 Functional Connectomes Database.," *Frontiers in human neuroscience,* vol. 6, p. 301, 1 3 2012.

[72] B. B. Biswal, M. Mennes, X. N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe, A. M. Dogonowski, M. Ernst, D. Fair, M. Hampson, M. J. Hoptman, J. S. Hyde, V. J. Kiviniemi, R. Kötter, S. J. Li, C. P. Lin and M. P. Milham, "Toward discovery science of human brain function," in *National Academy of Sciences of the United States of America*.

[73] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey and M. W. Weiner, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods.," *Journal of Magnetic Resonance Imaging,* vol. 27, no. 4, pp. 685-691, 2008.

[74] I. Beheshti and H. Demirel, "Feature-ranking-based Alzheimer's disease classification from structural MRI ," *Magnetic resonance imaging,* vol. 34(3), 2016.

[75] E. E. Bron, M. Smits, W. J. Niessen and S. Klein, "Feature Selection Based on the SVM Weight Vector for Classification of Dementia," *IEEE Journal of Biomedical and Health Informatics* , vol. 19 (5), p. 1617–1626, 1 9 2015.

[76] H.-I. Suk and D. Shen, "Clustering-induced multi-task learning for AD/MCI classification," *Medical Image Computing and Computer-assisted Intervention,* vol. 17(3), p. 393–400, 1 1 2014.

[77] M. Liu, D. Zhang, E. Adeli and D. Shen, "Inherent Structure-Based Multiview Learning With Multitemplate Feature Representation for Alzheimer's Disease Diagnosis," *IEEE Transactions on Bio-Medical Engineering,* vol. 63 (7), p. 1473–1482, 1 7 2016.

[78] L. Snoek, M. van der Miesen, T. Beemsterboer, A. van der Leij, A. Eigenhuis and H. Scholte, "The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses," *Sci Data,* vol. 8, no. 85, 2021.

[79] L. Alexander, J. Escalera and L. Ai, "An open resource for transdiagnostic research in pediatric mental health and learning disorders," *Sci Data 4,* 2017.

[80] T. He, R. Kong, A. J. Holmes, M. Nguyen, M. R. Sabuncu, S. B. Eickhoff, D. Bzdok, J. Feng and B. Yeo, "Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics," *NeuroImage,* vol. 206, p. 116276, 2020.

[81] B. Casey, T. Cannonier, M. I. Conley, A. O. Cohen, D. M. Barch, M. M. Heitzeg, M. E. Soules, T. Teslovich, D. V. Dellarco, H. Garavan, C. A. Orr, T. D. Wager, M. T. Banich, N. K. Speer, M. T. Sutherland, M. C. Riedel, A. S. Dick, J. M. Bjork, K. M. Thomas, B. Chaarani, M. H. Mejia, D. J. Hagler, M. D. Cornejo, C. S. Sicat, M. P. Harms, N. U. Dosenbach, M. Rosenberg, E. Earl, H. Bartsch, R. Watts, J. R. Polimeni, J. M. Kuperman, D. A. Fair and A. M. Dale, "The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites.," *Dev Cogn Neurosci,* vol. 32, pp. 43-54, 2018.

[82] C. L. Chen, D. B. Kaber and P. G. Dempsey, "A new approach to applying feedforward neural networks to the prediction of musculoskeletal disorder risk. ," *Applied ergonomics,* vol. 31(3), p. 269–282, 2000.

[83] I. Arel, D. C. Rose and T. P. Karnowski, "Deep machine learning- a new frontier in artificial intelligence research," *Computational Intelligence Magazine, IEEE,* vol. 5, no. 4, p. 13–18, 2010.

[84] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, p. 2278–2324, 1998.

[85] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao and Z. Liu, "Neural encoding and decoding with deep learning for dynamic natural vision," *Cerebral cortex (New York, N.Y. : 1991),* p. 1–25, 2017.

[86] Y. Zhang, T. Lee and M. Li, "Convolutional neural network models of V1 responses to complex patterns," *J Comput Neurosci,* vol. 46, p. 33–54, 2019.

[87] C. Zhang, K. Qiao, L. Wang, l. Tong, Y. Zeng and B. Yan, "Constraint-Free Natural Image Reconstruction From fMRI Signals Based on Convolutional Neural Network," *Frontiers in Human Neuroscience,* 2018.

[88] M. Svanera, M. Savardi, S. Benini, A. Signoroni, G. Raz, T. Hendler, L. Muckli, R. Goebel and G. Valente, "Transfer learning of deep neural network representations for fMRI decoding ," *Journal of neuroscience methods,* vol. 328, p. 108319, 2019.

[89] M. N. I. Qureshi, J. Oh and B. Lee, "3D-CNN based Discrimination of Schizophrenia using Resting-state fMRI," *Artificial Intelligence in Medicine,* vol. 98, 6 2019.

[90] P. Baldi, "Autoencoders, Unsupervised Learning, and Deep Architectures," in *ICML Unsupervised and Transfer Learning,* 2012.

[91] X. Han, Y. Zhong, L. He, S. Philip and L. Zhang, "The unsupervised hierarchical convolutional sparse auto-encoder for neuroimaging data classification," *International Conference on Brain Informatics and Health,* p. 156–166, 2015.

[92] H. Suk, S. Lee and D. Shen, "Alzheimer's disease neuroimaging initiative. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis," *Brain Struct. Funct.,* vol. 220, p. 841–859, 2015.

[93] F. Li, L. Tran, K. Thung, S. Ji, D. Shen and J. Li, "A Robust Deep Model for Improved Classification of AD/MCI Patients.," *IEEE Journal of Biomedical and Health Informatics. ,* 2015.

[94] P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning,* 2008.

[95] J. Kim, V. D. Calhoun, E. Shim and J. H. Lee, "Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. ," *NeuroImage,* vol. 124(Pt A), p. 127–146, 2016.

[96] 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, p. 16–23, 2017.

[97] X. Guo, K. C. Dominick, A. A. Minai, H. Li, C. A. Erickson and L. J. (. Lu, "Diagnosing Autism Spectrum Disorder from Brain Resting-State Functional Connectivity Patterns Using a Deep Neural Network with a Novel Feature Selection Method.," *Frontiers in neuroscience*, vol. 11, p. 460, 2017.

[98] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P.-A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local de- noising criterion.," *J. Mach. Learn. Res.* , vol. 11(3), p. 3371–3408, 2010.

[99] M. Sundararajan, A. Taly and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[100] S. Lapuschkin, A. Binder, G.́. Montavon, K.-R. Mu ̈ller and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[101] G.́. Montavon, S. Bach, A. Binder, W. Samek and K.-R. Mu ̈ller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, p. 211–222, 2017.

[102] A. Shrikumar, P. Greenside and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning*.

[103] G. Montavon, A. Binder, S. Lapuschkin, W. Samek and K.-R. Müller, " Layer-Wise Relevance Propagation: An Overview," 2019.

[104] V. Hangya, F. Braune, A. Fraser and H. Schütze, "Two Methods for Domain Adaptation of Bilingual Tasks: Delightfully Simple and Broadly Applicable.," in *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[105] E. Tzeng, J. Hoffman, K. Saenko and T. Darrell, "Adversarial discriminative domain adaptation," in *the IEEE Conference on Computer Vision and Pattern Recognition* , 2017.

[106] K. Kaisar, S. Mostafa, V. Sergi, R. Àlex, S. Joaquim, O. Arnau and L. Xavier, "Transductive Transfer Learning for Domain Adaptation in Brain Magnetic Resonance Image Segmentation," *Frontiers in Neuroscience*, vol. 15, p. 444, 2021.

[107] D. H. Geschwind and P. Levitt, "Autism spectrum disorders: developmental disconnection syndromes," *Current opinion in neurobiology*, vol. 17, p. 103–111, 2007.

[108] R. M. Jones and C. Lord, "Diagnosing autism in neurobiological research studies ," *Behavioural brain research* , vol. 251, p. 113–124.

[109] G. Dichter, "Functional magnetic resonance imaging of autism spectrum disorders ," *Dialogues in Clinical Neuroscience*, vol. 14, no. 3, p. 319–351, 2012.

[110] Y. Bengio, "Learning Deep Architectures for AI," *Foundations*, vol. 2, pp. 1-55, 2009.

[111] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Netw*, vol. 61, pp. 85-117, 2015.

[112] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677-691, 2017.

[113] D. Durstewitz, G. Koppe and A. Meyer-Lindenberg, "Deep neural networks in psychiatry," *Mol Psychiatry*, vol. 24, p. 1583–1598, 2019.

[114] S. M. G. Vieira, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neuroscience & Biobehavioral Reviews,* vol. 74, pp. 58-75, 2017.

[115] P. Lanka, D. Rangaprakash, M. Dretsch, J. Katz, T. Denney and G. Deshpande, "Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets," *Brain Imaging Behav.,* vol. 14, no. 6, pp. 2378-2416, 12 2020.

[116] R. Caruana, S. Lawrence and C. Giles, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping.," *Advances in Neural Information Processing Systems,* vol. 13, pp. 402-408, 2000.

[117] B. Mwangi, T. S. Tian and J. C. Soares, "A review of feature reduction techniques in neuroimaging.," *Neuroinformatics,* vol. 12, no. 2, p. 229–244, 2014.

[118] J. A. Nielsen, "Multisite functional connectivity MRI classification of autism: ABIDE results," *Front. Hum. Neurosci,* vol. 7, p. 599, 2013.

[119] A. Di Martino, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry* , vol. 19, p. 659–667, 2014.

[120] J. S. Anderson, J. A. Nielsen, A. L. Froehlich, M. B. Dubray, T. J. Druzgal and A. N. Cariello, "Functional connectivity magnetic resonance imaging classification of autism," *Brain* , vol. 134, p. 3742–3754, 2011.

[121] J. Yosinski, J. Clune, Y. Bengio and H. Lipson, "How transferable are features in deep neural networks?," in *In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14).* , 2014.

[122] Sarinnapakorn, Kanoksri and M. Kubat, "Combining Subclassifiers in Text Categorization: A DST-Based Solution and a Case Study," *IEEE Transactions on Knowledge and Data Engineering 19* , pp. 1638-1651, 2007.

[123] J. Zhang, W. Li, P. Ogunbona and D. Xu, "Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective," *ACM Comput. Surv.,* vol. 52, pp. 1-38, 2019.

[124] K. Miller, F. Alfaro-Almagro, N. Bangerter, D. Thomas, E. Yacoub, J. Xu, A. Bartsch, S. Jbabdi, S. Sotiropoulos, J. Andersson, L. Griffanti, G. Douaud, T. Okell, P. Weale, I. Dragonu, S. Garratt, S. Hudson, R. Collins, M. Jenkinson, P. Matthews and S. Smith, "Multimodal population brain imaging in the UK Biobank prospective epidemiological study.," *Nature Neuroscience,* vol. 19, no. 11, pp. 1523-1536, 2016.

[125] D. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. Curtiss, P. S. Della, D. Feinberg, M. Glasser, N. Harel, A. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, Petersen, S.E., F. Prior, B. Schlaggar, S. Smith, A. Snyder, J. Xu and E. Yacoub, "The Human Connectome Project: a data acquisition perspective.," *Neuroimage,* vol. 62, p. 2222–2231, 2012.

[126] L. Snoek, M. Miesen, T. Beemsterboer, A. Leij and H. Scholte, "The Amsterdam Open MRI Collection (AOMIC): A Collection of Publicly Available Population Imaging Datasets," *Sci Data 8,* vol. 85, 2021 url: https://doi.org/10.1038/s41597-021-00870-6.

[127] C. Doersch, "Tutorial on Variational Autoencoders," *ArXiv,* 2016 doi: https://arxiv.org/abs/1606.05908.

[128] S. Semeniuta, A. Severyn and E. Barth, "A Hybrid Convolutional Variational Autoencoder for Text Generation," pp. 627-637 , 2017.

[129] L. Zou, J. Zheng, C. Miao, M. McKeown and Z. Wang, "3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI," *IEEE Access,* vol. 5, pp. 23626-23636, 2017.

[130] E. Hosseini-Asl, M. Ghazal, A. Mahmoud, A. Aslantas, A. Shalaby, M. Casanova, G. Barnes, G. Gimel'farb, R. Keynton and A. El-Baz, "Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network," *Frontiers in bioscience,* vol. 23, pp. 584-596, 2018.

[131] B. Cheng, M. Liu, D. Zhang and D. Shen, "Robust multi-label transfer feature learning for early diagnosis of Alzheimer's disease," *Brain imaging and behavior,* vol. 13(1), p. 138–153, 2019.

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

[133] P. Vakli, R. J. Deák-Meszlényi, P. Hermann and Z. Vidnyánszky, "Transfer learning improves resting-state functional connectivity pattern analysis using convolutional neural networks," *GigaScience,* vol. 7, no. 12, p. 130, 2018.

[134] S. Smithetal, "Resting-state fMRI in the Human Connectome Project," *Neuroimage,* vol. 80, p. 144–168, 2013.

[135] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. Singh Khundrakpam, J. David Lewis, Q. Li, M. Milham, C. Yan and P. Bellec, "The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives.," *Frontiers in Neuroinformatics,* vol. 7, 2013 url: https://www.frontiersin.org/10.3389/conf.fninf.2013.09.00041.

[136] L. M. Alexander, J. Escalera, L. Ai, C. Andreotti, K. Febre, A. Mangone and M. P. (. Milham, "An open resource for transdiagnostic research in pediatric mental health and learning disorders.," *Scientific data,* vol. 4, no. 1, pp. 1-26, 2017.

[137] M. Martyn, "Testable Hypotheses for Unbalanced Neuroimaging Data," *Frontiers in Neuroscience,* vol. 10, p. 270, 2016.

[138] A. Estabrooks, D. T. Jo and N. Japkowicz, "A Multiple Resampling Method for Learning from Imbalanced Data Sets.," *Computational Intelligence. ,* vol. 20, p. 18–36, 2004.

[139] L. Yuan, Y. Wang, P. Thompson, V. Narayan and J. Ye, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data.," *Neuroimage,* vol. 61, no. 3, p. 622–632, 2012.

[140] H. He and E. Garcia, "Learning from Imbalanced Data.," *Knowledge and Data Engineering, IEEE Transactions on.,* vol. 21, no. 9, p. 1263–1284, 2009.

[141] H. Guo, Y. Li, J. Shang and M. Gu, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications,* vol. 73, pp. 220-239, 2017.

[142] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique.," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[143] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics,* vol. 14, p. 106, 2013.

[144] R. Blagus and L. Lusa, "Evaluation of SMOTE for High-Dimensional Class-Imbalanced Microarray Data," in *11th International Conference on Machine Learning and Applications,* 2012.

[145] D. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint,* 2014 arXiv:1312.6114..

[146] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair and A. B. Y. Courville, "Generative adversarial nets," *Advances in neural information processing systems,* vol. 27, pp. 2672-2680, 2014.

[147] M. Martín, B. Carro, A. Sánchez-Esguevillas and J. Lloret, "Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection," *IOT Sensors,* vol. 17, no. 9, p. 1967, 2017.

[148] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas and A. Malossi, "BAGAN: Data Augmentation with Balancing GAN," *ArXiv preprint ,* 2018 arXiv:1803.09655..

[149] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks, „," *Expert Systems with Applications,* vol. 91, pp. 464-471, 2018.

[150] P. Shamsolmoali, M. Zareapoor, L. Shen, A. H. Sadka and J. Yang, "Imbalanced data learning by minority class augmentation using capsule adversarial networks," *Neurocomputing,* vol. 459, pp. 481-493, 2021.

[151] V. A. Fajardo, D. Findlay, C. Jaiswal, X. Yin, R. Houmanfar, H. Xie, J. Liang, X. She and D. Emerson, "On oversampling imbalanced data with deep conditional generative models," *Expert Systems with Applications,* vol. 169, p. 114463, 2021.

[152] Q. Wang, F. Meng and T. Breckon, "Data Augmentation with norm-VAE for Unsupervised Domain Adaptation," *ArXiv preprint,* 2020 arXiv:2012.00848.

[153] K. Han, H. Wen, J. Shi, K.-H. Lu, Y. Zhang and Z. Liu, "Variational Autoencoder: An Unsupervised Model for Modeling and Decoding fMRI Activity in Visual Cortex," *NeuroImage,* vol. 198, pp. 125-136, 2019.

[154] J. Kawahara, C. Brown, S. Miller, B. Booth, V. Chau, R. Grunau, J. Zwicker and G. Hamarneh, "BrainNetCNN: Convolutional Neural Networks for Brain Networks; Towards Predicting Neurodevelopment," *NeuroImage,* vol. 146, pp. 1038-1049, 2017.

[155] R. Meszlényi, K. Buza and Z. Vidnyánszky, "Resting State fMRI Functional Connectivity-Based Classification Using a Convolutional Neural Network Architecture," *Frontiers in Neuroinformatics,* vol. 11, p. 61, 2017.

[156] X. Han, Y. Zhong, L. He, S. Philip and L. Zhang, "The unsupervised hierarchical convolutional sparse auto-encoder for neuroimaging data classification," in *International Conference on Brain Informatics and Health,* 2015.

[157] F. Li, L. Tran, K. H. Thung, S. Ji, D. Shen and J. Li, "A Robust Deep Model for Improved Classification of AD/MCI Patients ," *IEEE journal of biomedical and health informatics,* vol. 19, no. 5, pp. 1610-1616, 2015.

[158] Y. Li, J. Liu, J. Huang, Z. Li and P. Liang, "Learning Brain Connectivity Sub-networks by Group- Constrained Sparse Inverse Covariance Estimation for Alzheimer's Disease," *Frontiers in neuroinformatics,* vol. 12, p. 58, 2018.

[159] X. Chen, j. He, R. Lawrence and J. Carbonell, "Adaptive Multi-task Sparse Learning with an Application to fMRI Study," pp. 212-223, 2018 doi:10.1137/1.9781611972825.19.

[160] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B,* vol. 58, pp. 267-288, 1994.

[161] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne and P.-J. Kindermans, "iNNvestigate neural networks!," *J. Mach. Learn. Res.,* vol. 20, no. 93, pp. 1-8, 2018.

[162] K. Simonyan, A. Vedaldi and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv preprint,* 2014 arXiv:1312.6034.

[163] Z. Sherkatghanad, M. Akhondzadeh, S. Salari, M. Zomorodi-Moghadam, M. Abdar, U. R. Acharya, R. Khosrowabadi and V. Salari, " Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network," *Frontiers in neuroscience,* vol. 13, p. 1325, 2020.

[164] Y. Bai, Z. Pascal, W. Hu, V. D. Calhoun and Y. P. Wang, "Biomarker Identification Through Integrating fMRI and Epigenetics," *IEEE transactions on bio-medical engineering,* vol. 67, no. 4, p. 1186–1196, 2020.

[165] L. Kohoutová, J. Heo and S. Cha, "Toward a unified framework for interpreting machine-learning models in neuroimaging.," *Nat Protoc ,* vol. 15, p. 1399–1435 , 2020.

[166] A. Thomas, H. Heekeren, K. Müller and W. Samek, "Analyzing Neuroimaging Data Through Recurrent Deep Learning Models ," *Frontiers in Neuroscience, 13,* 2019 arXiv:1810.09945.

[167] Y. Chao-Gan and Z. Yu-Feng, "DPARSF: A MATLAB Toolbox for "Pipeline" Data Analysis of Resting-State fMRI," *Frontiers in systems neuroscience,* vol. 4, p. 13, 2010.

[168] L. He, H. Li, S. K. Holland, W. Yuan, M. Altaye and N. A. Parikh, "Early prediction of cognitive deficits in very preterm infants using functional connectome data in an artificial neural network framework," *Neuroimage Clin.,* vol. 18, p. 290–297, 2018.

[169] K. Ugurbil and J. Xu, "Pushing spatial and temporal resolution for functional and diffusion MRI in the human connectome project," *NeuroImage,* vol. 80 , p. 80–104 , 2013.

[170] S. Smith, M. Jenkinson, M. Woolrich, C. Beckmann, T. Behrens, H. Johansen-Berg, P. Bannister, M. De Luca, I. Drobnjak and D. Flitney, "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage,* vol. 23, pp. S208-S209, 2004.

[171] R. Craddock, G. James, P. Holtzheimer, X. Hu and H. Mayberg, "A whole brain fMRI atlas generated via spatially constrained spectral clustering ," *Human brain mapping,* vol. 33, no. 8, pp. 1914-28, 2012.

[172] D. Handwerker, J. Ollinger and M. D'Esposito, "Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses," *Neuroimage.,* vol. 21, no. 4, pp. 1639-51, 4 2004.

[173] G. Aguirre, E. Zarahn and M. D'Esposito, "The variability of human BOLD hemodynamic responses.," *NeuroImage,* vol. 8, no. 4, pp. 360-369, 1998.

[174] D. Rangaprakash, G.-R. Wu, D. Marinazzo, X. Hu and G. Deshpande, "Hemodynamic response function (HRF) variability confounds resting state fMRI connectivity," *Magnetic Resonance in Medicine,* vol. 80, no. 4, pp. 1697-1713, 2018.

[175] D. Rangaprakash, G.-R. Wu, D. Marinazzo, X. Hu and G. Deshpande, "Parameterized hemodynamic response function data of healthy individuals obtained from resting-state functional MRI in a 7T MRI scanner," *Data in Brief.,* vol. 17, pp. 1175-1179, 2018.

[176] D. Rangaprakash, M. Dretsch, W. Yan, J. Katz, T. Denney and G. Deshpande, "Hemodynamic variability in soldiers with trauma: Implications for functional MRI connectivity studies," *NeuroImage: Clinical,* vol. 16, pp. 409-417, 2017.

[177] W. Yan, L. Palaniyappan, P. Liddle, R. Deshpande and G. Deshpande, "Characterizations of hemodynamic alterations in Schizophrenia and Bipolar disorder and their effect on resting-state fMRI functional connectivity," *Schizophrenia Bulletin,* 2021(in press).

[178] W. Yan, R. Deshpande and G. Deshpande, "Aberrant hemodynamic responses in Autism: Implications for resting state fMRI functional connectivity studies," *NeuroImage: Clinical,* vol. 19, pp. 320-330, 2018.

[179] W. Yan, R. Deshpande and D. Gopikrishna, "Hemodynamic response function parameters obtained from resting state BOLD fMRI data in subjects with Autism Spectrum Disorder and matched healthy controls," *Data in Brief,* vol. 14, pp. 558-562, 2017.

[180] M. Boly, S. Sasai, O. Gosseries, M. Oizumi, A. Casali, M. Massimini and G. Tononi, "Stimulus set meaningfulness and neurophysiological differentiation: a functional magnetic resonance imaging study.," *PloS one,* vol. 10, no. 5, 2015 doi: 10.1371/journal.pone.0125337.

[181] D. Rangaprakash, R. Tadayonnejad, G. Deshpande, J. O'Neill and J. D. Feusner, "FMRI hemodynamic response function (HRF) as a novel marker of brain function: applications for understanding obsessive-compulsive disorder pathology and treatment response.," *Brain imaging and behavior,* vol. 15, no. 3, p. 1622–1640, 2021.

[182] E. Amico, F. Gomez, C. Di Perri, A. Vanhaudenhuyse, D. Lesenfants, P. Boveroux, V. Bonhomme, J. F. Brichant, D. Marinazzo and S. Laureys, "Posterior cingulate cortex-related co-activation patterns: a resting state FMRI study in propofol-induced loss of consciousness.," *PloS one,* vol. 9, no. 6, p. e100012, 2014.

[183] D. Rangaprakash, G. Deshpande, T. A. Daniel, A. M. Goodman, J. L. Robinson, N. Salibi, J. S. Katz, T. S. J. Denney and M. N. Dretsch, "Compromised hippocampus-striatum pathway as a potential imaging biomarker of mild-traumatic brain injury and posttraumatic stress disorder.," *Human brain mapping,* vol. 38, no. 6, p. 2843–2864, 2017.

[184] B. Lamichhane, B. M. Adhikari, S. F. Brosnan and M. Dhamala, "The neural basis of perceived unfairness in economic exchanges.," *Brain connectivity,* vol. 4, no. 8, p. 619–630, 2014.

[185] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint,* 2014 arXiv:1412.6980..

[186] W. Shi, J. Caballero, L. Theis, F. Huszar, A. Aitken, C. Ledig and Z. Wang, "Is the deconvolution layer the same as a convolutional layer?," *arXiv preprint*, 2016 arXiv:1609.07009..

[187] M. D. Zeiler, D. Krishnan, G. W. Taylor and R. Fergus, "Deconvolutional networks," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010.

[188] D. Kingma, D. Rezende, S. Mohamed and M. Welling, "Semi-Supervised Learning with Deep Generative Models," in *Advances in Neural Information Processing Systems*, 2014.

[189] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jo´zefowicz and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint* , p. 10–21, 2015 arXiv:1511.06349..

[190] J. Ji and Y. Yao, "A novel CNN framework to extract multi-level modular features for the classification of brain networks," *Applied Intelligence*, 2021 doi: https://doi.org/10.1007/s10489-021-02668-w.

[191] J.-H. Kim, Y. Zhang, K. Han, Z. Wen, M. Choi and Z. Liu, "Representation Learning of Resting State fMRI with Variational Autoencoder," *NeuroImage*, vol. 241, p. 118423, 2021.

[192] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint*, 2018 arXiv:1803.08375..

[193] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res*, vol. 15, p. 1929–1958, 2014.

[194] S. Wager, S. Wang and P. S. Liang, "Dropout Training as Adaptive Regularization," *Advances in Neural Information Processing Systems 26*, p. 351–359, 2013.

[195] H. Sewani and R. Kashef, "An Autoencoder-Based Deep Learning Classifier for Efficient Diagnosis of Autism," *Children (Basel, Switzerland)*, vol. 7, no. 10, p. 182, 2020.

[196] 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," *Frontiers in Neuroinformatics*, vol. 13, p. 70, 2019.

[197] W. Yin, S. Mostafa and F.-x. Wu, "Diagnosis of Autism Spectrum Disorder Based on Functional Brain Networks with Deep Learning," *Journal of Computational Biology*, vol. 28, no. 2, p. 146, 2021.

[198] K. H. Brodersen, C. S. Ong, S. K. E. and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in *2010 20th International Conference on Pattern Recognition*, Istanbul, 2010.

[199] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, C. G.S., A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems.," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, 2016.

[200] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, p. 8024–8035.

[201] A. Binder, S. Lapuschkin, G. Montavon, K.-R. Müller and W. Samek, "Layer-wise relevance propagation for deep neural network architectures.," *In Information science and applications (ICISA)*, pp. 913-922, 2016.

[202] M. Böhle, F. Eitel, M. Weygandt and K. Ritter, "Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification," *Frontiers in aging neuroscience*, vol. 11, p. 194, 2019.

[203] A. Binder, G. Montavon, S. Lapuschkin, K. Müller and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers.," in *In International Conference on Artificial Neural Networks*, 2016.

[204] G. Montavon, S. Lapuschkin, A. Binder, W. Samek and K.-R. Müller, "Explaining NonLinear Classification Decisions with Deep Taylor Decomposition," *Pattern Recognition*, vol. 65, p. 211–222, 2017.

[205] G. Montavon, W. Samek and K. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process*, vol. 73, pp. 1-15, 2018.

[206] M. Xia, J. Wang and Y. He, "BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics," *PLoS ONE*, vol. 8, no. 7, p. e68910, 2013.

[207] N. Chaitra, P. Vijaya and G. Deshpande, "Diagnostic Prediction of Autism Spectrum Disorder Using Complex Network Measures in a Machine Learning Framework," *Biomedical Signal Processing and Control (in press)*, vol. 62, p. 102099, 2020.

[208] M. Syed, Z. Yang, X. Hu and G. Deshpande, "Investigating Brain Connectomic Alterations in Autism using the Reproducibility of Independent Components derived from Resting State functional MRI Data," *Frontiers in Neuroscience*, vol. 11, p. 459, 2017.

[209] G. Deshpande, L. Libero, K. Sreenivasan, H. Deshpande and R. Kana, "Identification of neural connectivity signatures of autism using machine learning," *Frontiers in Human Neuroscience*, vol. 7, p. 670, 2013.

[210] R. K. Kana, L. Q. Uddin, T. Kenet, C. Diane and M. Ralph-Axel, "Brain connectivity in autism," *Frontiers in Human Neuroscience*, vol. 8, p. 349, 2014.

[211] V. Cherkassky, R. Kana, T. Keller and M. Just, "Functional connectivity in a baseline resting-state network in autism," *NeuroReport*, vol. 17 (16), p. 1687–1690, 2006.

[212] E. A. von dem Hagen, R. S. Stoyanova, S. Baron-Cohen and A. J. Calder, "Reduced functional connectivity within and between 'social' resting state networks in autism spectrum conditions," *Social cognitive and affective neuroscience*, vol. 8(6), p. 694–701, 2013.

[213] H. Li, N. A. Parikh and L. He, "A Novel Transfer Learning Approach to Enhance Deep Neural Network Classification of Brain Functional Connectomes," *Frontiers in Neuroscience*, no. 12, p. 491, 2018.

[214] D. Karimi, H. Dou, S. K. Warfield and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, 2020.

[215] B.-D. Shai, B. John, C. Koby, K. Alex, P. Fernando and W. V. Jennifer, "A theory of learning from different domains.," *Machine learning*, vol. 79, p. 151–175, 2010.

[216] X. Li, Y. Gu, N. Dvornek, L. Staib, P. Ventola and J. Duncan, "Multi-site fMRI Analysis Using Privacy-preserving Federated Learning and Domain Adaptation: ABIDE Results," *Medical Image Analysis*, p. 101765 , 2020.

[217] S. Zhou, C. Cox and H. Lu, "Improving Whole-Brain Neural Decoding of fMRI with Domain Adaptation," *bioRxiv*, p. doi: https://doi.org/10.1101/375030, 2018.

[218] G. Csurka, "A comprehensive survey on domain adaptation for visual applications.," *Domain adaptation in computer vision applications* , pp. 1-35, 2017.

[219] B. Gholami, P. Sahu, O. Rudovic, K. Bousmalis and V. Pavlovic, "Unsupervised multi-target domain adaptation: An information theoretic approach," *IEEE Transactions on Image Processing*, vol. 29, pp. 3993-4002, 2020.

[220] J. Hoffman, M. Mohri and N. Zhang, "Algorithms and theory for multiple-source adaptation," *arXiv preprint*, p. doi: https://arxiv.org/abs/1805.08727, 2018.

[221] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai and K. Keutzer, "Multi-source domain adaptation for semantic segmentation," *arXiv preprint*, p. doi: https://arxiv.org/abs/1910.12181, 2019.

[222] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. de Leeuw, C. M. Tempany and B. van Gin- neken, "Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation," in *International conference on medical image computing and computer-assisted intervention*, 2017.

[223] C. Wachinger, M. Reuter and A. D. N. Initiative, "Domain adaptation for Alzheimer's disease diagnostics," *NeuroImage*, vol. 139, p. 470–479, 2016.

[224] S. Bickel and M. Brückner, "Discriminative Learning for Differing Training and Test Distributions," in *Proceedings of the 24th international conference on Machine learning*, 2007.

[225] N. Duc, S. Ryu and M. Qureshi, "3D-deeplearningbased automatic diagnosis of Alzheimer's disease with joint MMSE prediction using resting-state fMRI," *Neuroinformatics*, vol. 18, no. 1, pp. 71-86, 2020.

[226] M. Cao, M. Yang, C. Qin, X. Zhu, Y. Chen, J. Wang and T. Liu, "Using DeepGCN to identify the autism spectrum disorder from multi-site resting-state data," vol. 70, p. 103015, 2021.

[227] F. Subah, K. Deb, P. Dhar and T. Koshiba, "A Deep Learning Approach to Predict Autism Spectrum Disorder Using Multisite Resting-State fMRI.," *Appl. Sci.*, vol. 11, no. 8, p. 3636, 2021.

[228] B. Cheng, D. Zhang and D. Shen, "Domain transfer learning for MCI conversion prediction.," *Med Image Comput Comput Assist Interv.*, vol. 15, pp. 82-90, 2012.

[229] N. M. Khan, N. Abraham and M. Hon, "Transfer Learning With Intelligent Training Data Selection for Prediction of Alzheimer's Disease," *IEEE Access*, vol. 7, pp. 72726-72735, 2019.

[230] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint,* p. doi: https://arxiv.org/abs/1409.1556, 2015.

[231] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. v. d. Laak, B. v. Ginneken and C. I. Sánchez, "A survey on deep learning in medical image analysis.," *Med Image Anal,* vol. 42, p. 60–88, 12 2017.

[232] A. Choudhary, L. Tong, Y. Zhu and M. D. Wang, "Advancing Medical Imaging Informatics by Deep Learning-Based Domain Adaptation.," *Yearbook of medical informatics,* vol. 29, no. 1, pp. 129-138, 2020.

[233] H. Guan and M. Liu, "Domain Adaptation for Medical Image Analysis: A Survey," *arXiv preprint,* p. doi: https://arxiv.org/abs/2102.09508, 2021.

[234] A. Madani, M. Moradi, A. Karargyris and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation," in *2018 IEEE 15th International Symposium on Biomedical Imaging*, 2018.

[235] M. Belhaj, P. Protopapas and W. Pan, "Deep Variational Transfer: Transfer Learning through Semi-supervised Deep Generative Models.," *ArXiv,* p. doi: http://arxiv.org/abs/1812.03123, 2018.

[236] H. Chen and J. Chien, "Deep semi-supervised learning for domain adaptation," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)* , Boston, MA, 2015.

[237] D. Kingma, D. Rezende, S. Mohamed and M. Welling, "Semi-Supervised Learning with Deep Generative Models," *Advances in Neural Information Processing Systems,* vol. 4, pp. 3581-3589, 2014.

[238] C. Louizos, K. Swersky, Y. Li, M. Welling and R. Zemel, "The Variational Fair Autoencoder," *arXiv preprint,* p. doi: https://arxiv.org/abs/1511.00830, 2016.

[239] M. Ilse, J. Tomczak, C. Louizos and M. Welling, "DIVA: Domain Invariant Variational Autoencoders.," in *Medical Imaging with Deep Learning*, 2020.

[240] S. Purushotham, W. Carvalho, T. Nilanon and Y. Liu, "Variational Recurrent Adversarial Deep Domain Adaptation," *The International Conference on Learning Representations,* p. URL: https://openreview.net/forum?id=rk9eAFcxg., 2017.

[241] A. Gretton, K. Borgwardt, M. Rasch and B. l. A. Schölkopf, "A Kernel Two-Sample Test," *The Journal of Machine Learning Research,* vol. 13, pp. 723-773, 2012.

[242] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," in *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*, 2008.

[243] L. Alexander, J. Escalera and L. Ai, "An open resource for transdiagnostic research in pediatric mental health and learning disorders.," *Sci Data,* vol. 4, p. 170181, 2017.

[244] L. Snoek, M. van der Miesen, T. Beemsterboer, A. van der Leij, A. Eigenhuis and H. Scholte, "The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses," *Sci Data,* vol. 8, p. 85, 2021.

[245] W. E. Johnson, C. Li and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics,* vol. 8, no. 1, p. 118–127, 1 2007.

[246] J. P. Fortin, E. M. Sweeney, J. Muschelli, C. M. Crainiceanu, R. T. Shinohara and A. D. N. Initiative, "Removing inter-subject technical variability in magnetic resonance imaging studies," *NeuroImage,* vol. 132, p. 198–212, 2016.

[247] L. Nyúl, J. Udupa and X. Zhang, "New variants of a method of mri scale standardization," *IEEE Trans Med Imaging.,* vol. 19, no. 2, p. 143–50, 2 2000.

[248] J. P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, M. McInnis, M. L. Phillips, M. H. Trivedi, M. M. Weissman and R. T. Shinohara, "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage,* vol. 167, pp. 104-120, 2018.

[249] J. P. Fortin, D. Parker, B. Tunç, T. Watanabe, M. A. Elliott, K. Ruparel, D. R. Roalf, T. D. Satterthwaite and R. C. Gur, "Harmonization of multi-site diffusion tensor imaging data," *NeuroImage,* vol. 161, p. 149–170, 2017.

[250] M. Yu, K. A. Linn, P. A. Cook, M. L. Phillips, M. McInnis, M. Fava, M. H. Trivedi, M. M. Weissman, R. T. Shinohara and Y. I. Sheline, "Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data Hum," *Brain Mapp.,* vol. 39, pp. 4213-4227, 2018.

[251] F. Zhao, Z. Chen, I. Rekik, S. W. Lee and D. Shen, "Diagnosis of Autism Spectrum Disorder Using Central-Moment Features From Low- and High-Order Dynamic Resting-State Functional Connectivity Networks.," *Frontiers in neuroscience,* vol. 14, p. 258, 2020.

[252] Q. Gu, Z. Li and J. Han, "Joint feature selection and subspace learning," *International Joint Conference on Artificial Intelligence,* vol. 22, p. 1294, 2011.

[253] J. Li, J. Zhao and K. Lu, "Joint Feature Selection and Structure Preservation for Domain Adaptation," *In International Joint Conference on Artificial Intelligence,* p. 1697–1703, 2016.

[254] B. Schölkopf, J. Platt and T. Hofmann, "Analysis of representations for domain adaptation," in *Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, 2007.

[255] F. Sun, H. Wu, Z. Luo, W. Gu, Y. Yan and Q. Du, "Informative Feature Selection for Domain Adaptation," *IEEE Access* , vol. 7, pp. 142551-142563, 2019.

[256] P. Li, Y. Ying and C. Campbell, "A variational approach to semi-supervised clustering," in *In Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2009.

[257] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmos. Environ.,* vol. 32, p. 2627–2636, 1998.

[258] J. Schmidhuber, "Deep learning in neural networks: an overview.," *Neural Netw.,* vol. 61, p. 85–117, 2015.

[259] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *The International Conference on Machine Learning,* p. 807–814, 2021.

[260] B. Mwangi, T. S. Tian and J. C. Soares, "A review of feature reduction techniques in neuroimaging.," *Neuroinformatics.,* vol. 12, no. 2, p. 229–244, 2014.

[261] T. Eslami, A. Fahad, J. S. Raiker and F. Saeed, "Machine Learning Methods for Diagnosing Autism Spectrum Disorder and Attention- Deficit/Hyperactivity Disorder Using Functional and Structural MRI: A Survey," *Frontiers in Neuroinformatics,* vol. 14, p. 62, 2021.

[262] X.-a. Bi, Y. Wang, Q. Shu, Q. Sun and Q. Xu, "Classification of Autism Spectrum Disorder Using Random Support Vector Machine Cluster," *Frontiers in genetics,* vol. 9, p. 18, 2018.

[263] H. Chen, X. Duan, F. Liu, F. Lu, X. Ma and Y. Zhang, "Multivariate classification of autism spectrum disorder using frequency-specific resting-state functional connectivity—a multi-center study.," *Prog. Neuro Psychopharmacol. Biol. Psychiatry,* vol. 64, pp. 1-9, 2016.

[264] G. Chanel, S. Pichon, L. Conty, S. Berthoz, C. Chevallier and J. Grèzes, "Classification of autistic individuals and controls using cross-task characterization of fMRI activity," *NeuroImage: Clinical,* vol. 10, pp. 78-88, 2016.

[265] 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, p. 16–23, 2017.

[266] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. S. Khundrakpam, J. D. Lewis, Q. Li, M. Michael, C. Yan and P. Bellec, "The Neuro Bureau Preprocessing Initiative: open sharing of preprocessed neuroimaging data and derivatives.," *Frontiers in Neuroinformatics,* vol. 7, p. doi: 10.3389/conf.fninf.2013.09.00041, 2013.

[267] S. Whitfield-Gabrieli and A. Nieto-Castanon, "Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks," *Brain connectivity,* vol. 2, no. 3, p. 125–141, 2012.

[268] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko and T. Darrell, "Deep Domain Confusion: Maximizing for Domain Invariance. ," *ArXiv,* p. https://arxiv.org/abs/1412.3474, 2014.

[269] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research,* vol. 9, p. 2579–2605, 11 2008.

[270] V. Nozais, P. Boutinaud, V. Verrecchia, M. F. Gueye, P. Y. Hervé, C. Tzourio, B. Mazoyer and M. Joliot, "Deep Learning-based Classification of Resting-state fMRI Independent-component Analysis.," *Neuroinformatics,* pp. 1-19, 2021.

[271] D. Pedamonti, "Comparison of non-linear activation functions for deep neural networks on MNIST classification task," *ArXiv,* p. doi: https://arxiv.org/abs/1804.02763, 2018.

[272] J.-P. Fortin and W. Foran, "ComBatHarmonization," 2021. [Online]. Available: https://github.com/Jfortin1/ComBatHarmonization.

[273] F. Pedregosa, . Varoquaux and A. Gramfort, "Scikit-learn: Machine Learning in Python," *Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[274] P. Vakli, R. J. Deák-Meszlényi, P. Hermann and Z. Vidnyánszky, "Transfer learning improves resting-state functional connectivity pattern analysis using convolutional neural networks," *GigaScience,* vol. 7, no. 12, p. 130, 2018.

[275] 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.

[276] Y. Benjamini and Y. Hochberg, "Controling the false discovery rate: A practical and powerful approach to multiple testing.," *Journal of the Royal Statistical Society. Series B, Statistical Methodology,* vol. 57, pp. 289-300, 1995.

[277] W. M. Kouw and M. Loog, "A Review of Domain Adaptation without Target Labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 43, no. 3, pp. 766-785, 2021.

[278] F. H. Duffy, A. Shankardass, G. B. McAnulty and H. Als, "The relationship of Asperger's syndrome to autism: a preliminary EEG coherence study," *BMC medicine,* vol. 11, p. 175, 2013.

[279] F. Faridi and R. Khosrowabadi, "Behavioral, Cognitive and Neural Markers of Asperger Syndrome," *Basic and Clinical Neuroscience,* vol. 8, no. 5, pp. 349-360, 2017.

[280] J. Wang, L. Zhang, Q. Wang, L. Chen, J. Shi, X. Chen and Z. Li, "Multi-Class ASD Classification Based on Functional Connectivity and Functional Correlation Tensor via Multi-Source Domain Adaptation and Multi-View Sparse Representation.," *IEEE Transactions on Medical Imaging,* vol. 39, no. 10, pp. 3137-3147, 2020.

[281] M. Isam, N. Yahya, I. Faye and A. Faeq Hussein, "Identification of Autism Subtypes Based on Wavelet Coherence of BOLD FMRI Signals Using Convolutional Neural Network," *Sensors,* vol. 21, p. 5256, 2021.

[282] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert and B. Glocker, "Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks," *Information Processing In Medical Imaging.,* pp. 597-609, 2017.

[283] P. Haeusser, T. Frerix, A. Mordvintsev and D. Cremers, "Associative Domain Adaptation," *IEEE International Conference on Computer Vision,* pp. 2784-2792, 2017.

[284] S. R. Panta, R. Wang, J. Fries, R. Kalyanam, N. Speer, M. Banich, K. Kiehl, M. King, M. Milham, T. D. Wager, J. A. Turner, S. M. Plis and V. D. Calhoun, "A Tool for Interactive Data Visualization: Application to Over 10,000 Brain Imaging and Phantom MRI Data Sets," *Frontiers in neuroinformatics,* vol. 10, no. 9, p. doi: https://doi.org/10.3389/fninf.2016.00009, 2016.

[285] P. Washington, K. M. Paskov, H. Kalantarian, N. Stockham, C. Voss, A. Kline, R. Patnaik, B. Chrisman, M. Varma, Q. Tariq, K. Dunlap, J. Schwartz, N. Haber and D. P. Wall, "Feature Selection and Dimension Reduction of Social Autism Data. Pacific Symposium on Biocomputing," *Pacific Symposium on Biocomputing,* vol. 25, p. 707–718, 2020.

[286] C. Shi, X. Xin and J. Zhang, "Domain Adaptation Using a Three-Way Decision Improves the Identification of Autism Patients from Multisite fMRI Data," *Brain Sci.,* vol. 11, no. 5, p. 603, 2021.

[287] M. Wang, D. Zhang, J. Huang, P. Yap, D. Shen and M. Liu, "Identifying Autism Spectrum Disorder With Multi-Site fMRI via Low-Rank Domain Adaptation," *IEEE Transactions on Medical Imaging,* vol. 39, pp. 644-655, 2020.

[288] K. Miller, F. Alfaro-Almagro and N. Bangerter, "Multimodal population brain imaging in the UK Biobank prospective epidemiological study," *Nat. Neurosci,* vol. 19, p. 1523–1536, 2016.

[289] A. Holmes, M. Hollinshead and T. O'Keefe, "Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures.," *Sci Data 2,* p. 150031, 2015.

[290] K. B. Nooner, S. J. Colcombe, R. H. Tobe, M. Mennes, M. M. Benedict, A. L. Moreno, L. J. Panek, S. Brown, S. T. Zavitz, Q. Li, S. Sikka, D. Gutman, S. Bangaru, R. T. Schlachter, S. M. Kamiel, A. R. Anwar, C. M. Hinz, M. S. Kaplan, A. B. Rachlin and S. Adelsberg, "The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry.," *Frontiers in neuroscience,* vol. 6, p. 152, 2012.

[291] K. Kalcher, W. Huf, R. Boubela, P. Filzmoser, L. Pezawas, B. Biswal, S. Kasper, E. Moser and C. Windischberger, "Fully Exploratory Network Independent Component Analysis of the 1000 Functional Connectomes Database," *Frontiers in human neuroscience,* vol. 6, p. 301, 1 3 2012.

[292] T. D. Satterthwaite, J. J. Connolly, K. Ruparel, M. E. Calkins, C. Jackson, M. A. Elliott, D. R. Roalf, R. Hopson, K. Prabhakaran, M. Behr, H. Qiu, F. D. Mentch, R. Chiavacci, P. M. A. Sleiman, R. C. Gur, H. Hakonarson and R. E. Gur, "The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth.," *Neuroimage.(Pt B),* vol. 124, pp. 1115-1119, 2016.

[293] M. Rakić, M. Cabezas, K. Kushibar, A. Oliver and X. Lladó, "Improving the detection of autism spectrum disorder by combining structural and functional MRI information," *NeuroImage : Clinical,* vol. 25, p. 102181, 2020.

[294] M. Verly, J. Verhoeven, I. Zink, D. Mantini, L. Van Oudenhove, L. Lagae, S. Sunaert and N. Rommel, "Structural and functional underconnectivity as a negative predictor for language in autism. ," *Human brain mapping ,* vol. 35, no. 8, p. 3602–3615, 2014.

[295] H. Koshino, R. K. Kana, T. A. Keller, V. L. Cherkassky, N. J. Minshew and M. A. Just, "fMRI investigation of working memory for faces in autism: visual coding and underconnectivity with frontal areas.," *Cerebral Cortex,* vol. 18, p. 289–300, 2008.

[296] N. J. Minshew and T. A. Keller, "The nature of brain dysfunction in autism: functional brain imaging studies.," *Current opinion in neurology,* vol. 23, no. 2, p. 124–130, 2010.

[297] P. C. Pantelis, L. Byrge, J. M. Tyszka, R. Adolphs and D. P. Kennedy, "A specific hypoactivation of right temporo-parietal junction/posterior superior temporal sulcus in response to socially awkward situations in autism. ," *Social cognitive and affective neuroscience,* vol. 10, no. 10, p. 1348–1356, 2015.

[298] T. P. DeRamus, B. S. Black, M. R. Pennick and R. K. Kana, "Enhanced parietal cortex activation during location detection in children with autism.," *Journal of neurodevelopmental disorders,* vol. 6, no. 1, p. 37, 2014.

[299] S. Xu, M. Li and C. Yang, "Altered Functional Connectivity in Children With Low-Function Autism Spectrum Disorders," *Front Neurosci.,* vol. 13, p. 806, 2019.

[300] B. Horwitz, J. Rumsey, C. Grady and S. Rapoport, "The cerebral metabolic landscape in autism. Intercorrelations of regional glucose utilization.," *Arch Neurol,* vol. 45, no. 7, p. 749–755, 1988.

[301] G. S. Dichter, "Functional magnetic resonance imaging of autism spectrum disorders," *Dialogues in clinical neuroscience,* vol. 14, no. 3, p. 319–351, 2012.

[302] E. Hagen, R. Stoyanova, S. Baron-Cohen and A. Calder, "Reduced functional connectivity within and between 'social' resting state networks in autism spectrum conditions," *Social cognitive and affective neuroscience,* vol. 8, no. 6, pp. 694-701, 2013.

[303] J. Lee, E. Bigler and A. Alexander, "Diffusion tensor imaging of white matter in the superior temporal gyrus and temporal stem in autism," *Neuroscience Letters,* vol. 424, p. 127–32, 2007.

[304] D. Yan, S. Wu, M. T. Sami, A. Almudaifer, Z. Jiang, H. Chen, D. Rangaprakash, G. Deshpande and Y. Ma, "Improving Brain Dysfunction Prediction by GAN: A Functional-Connectivity Generator Approach," in *2021 IEEE International Conference on Big Data (Big Data),* 2021.

[305] T. Eslami, M. Vahid, F. Alvis, L. Angela and F. Saeed, "ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data," *Frontiers in Neuroinformatics,* vol. 13, p. 70, 2019.

[306] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Research,* vol. 16, pp. 321-357, 2002.

[307] F. Mahmood, R. Chen and N. J. Durr, "Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training," *IEEE Transactions on Medical Imaging,* vol. 37, no. 12, pp. 2572-2581, 12 2018.

[308] M. Greicius, B. Krasnow, A. Reiss and V. Menon, "Functional connectivity in the resting brain: A network analysis of the default mode hypothesis," in *Proceedings of the National Academy of Sciences of the United States of America* , 01 02 2003.

[309] Y. Bengio, "Learning Deep Architectures for AI," *Foundations, vol. 2, pp. 1-55, 2009.*.

[310] D. H. Geschwind and P. Levitt, "Autism spectrum disorders: developmental disconnection syndromes ," in *Current opinion in neurobiology*, 2007.

[311] G. . Dichter, "Functional magnetic resonance imaging of autism spectrum disorders ," *Dialogues in Clinical Neuroscience,* vol. 14(3), p. 319–351, 2012.

[312] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," pp. 2625-2634., 2015.

[313] P. Lanka, D. Rangaprakash, M. Dretsch, J. Katz, T. Denney and G. Deshpande, "Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets," *Brain Imaging and Behavior* , 2019 (in press).

[314] J. A. Nielsen, "Multisite functional connectivity MRI classification of autism: ABIDE results," *Front. Hum. Neurosci*, 2013.

[315] K. Simonyan, A. Vedaldi and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *CoRR* , 2014.

[316] F. Li, L. Tran, K. Thung, S. Ji, D. Shen and J. Li, "A Robust Deep Model for Improved Classification of AD/MCI Patients ," *IEEE J Biomed Health Inform,* vol. 19(5), pp. 1610-1616, 2015.

[317] J. Zhang, W. Li, P. Ogunbona and D. Xu, "Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective," *ACM Comput. Surv.,* vol. 52, pp. 7:1-7:38.

[318] D. Kingma and M. Welling, "Auto-Encoding Variational Bayes," 2014.

[319] K. Han, H. Wen, J. Shi, K.-H. Lu, Y. Zhang and Z. Liu, "Variational Autoencoder: An Unsupervised Model for Modeling and Decoding fMRI Activity in Visual Cortex," 2017.

[320] C. Doersch, "Tutorial on variational autoencoders," 2016.

[321] P. Vakli, R. J. Deák-Meszlényi, P. Hermann and Z. Vidnyánszky, "Transfer learning improves resting-state functional connectivity pattern analysis using convolutional neural networks," *GigaScience,* vol. 7(12) , p. giy130, 2018.

[322] A. Thomas, H. Heekeren, K. Müller and W. Samek, "Analyzing Neuroimaging Data Through Recurrent Deep Learning Models ," *Frontiers in Neuroscience, 13*, 2019.

[323] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. Singh Khundrakpam, J. David Lewis, Q. Li, M. Milham, C. Yan and P. Bellec, "The Neuro Bureau Preprocessing Initiative: open sharing of preprocessed neuroimaging data and derivatives," *In Neuroinformatics 2013, Stockholm, Sweden* , 2013.

[324] L. He, H. Li, S. K. Holland, W. Yuan, M. Altaye and N. A. Parikh, "Early prediction of cognitive deficits in very preterm infants using functional connectome data in an artificial neural network framework," *Neuroimage Clin.,* Vols. 18,, p. 290–297, 2018.

[325] Glasser, "The minimal preprocessing pipelines for the Human Connectome Project," 2013.

[326] S. Smith, M. Jenkinson, M. Woolrich, C. Beckmann, T. Behrens, H. Johansen-Berg, P. Bannister, M. De Luca, I. Drobnjak and D. Flitney, "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage,* vol. 23, p. S208, 2004.

[327] F. H. Van Eemeran and R. Grootendorst, "A Systematic Theory of Argumentation," in *Philosophy: 12.* , 2004.

[328] W. Shi, J. Caballero, L. Theis, F. Huszar, A. Aitken, C. Ledig and Z. Wang, "Is the deconvolution layer the same as a convolutional layer?," 2016.

[329] D. Kingma, D. Rezende, S. Mohamed and M. Welling, "Semi-Supervised Learning with Deep Generative Models," *Advances in Neural Information Processing Systems,* vol. 4, 2014.

[330] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jo´zefowicz and S. Bengio, "Generating sentences from a continuous space," *CONLL,* p. 10–21, 2016.

[331] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014.

[332] J. Kawahara, C. Brown, S. Miller, B. Booth, V. Chau, R. Grunau, J. Zwicker and G. Hamarneh, "BrainNetCNN: Convolutional Neural Networks for Brain Networks; Towards Predicting Neurodevelopment," *NeuroImage. 146.* , 2016.

[333] A. Binder, S. Lapuschkin, G. Montavon, K.-R. Müller and W. Samek, "Layer-Wise Relevance Propagation for Deep Neural Network Architectures," 2016.

[334] A. Binder, G. Montavon, S. Lapuschkin, K. Müller and W. Samek, "Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers," 2016.

[335] K. H. Brodersen, C. S. Ong, S. K. E. and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in *2010 20th International Conference on Pattern Recognition* , Istanbul, 2010.

[336] M. Xia, J. Wang and Y. He, "BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics," *PLoS ONE,* vol. 8(7), p. e68910, 2013.

[337] H. Li, N. Parikh and L. He, "A Novel Transfer Learning Approach to Enhance Deep Neural Network Classification of Brain Functional Connectomes. ," *Front. Neurosci.,* 2018.

[338] E. Hagen, R. Stoyanova, S. Baron-Cohen and A. Calder, "Reduced functional connectivity within and between 'social' resting state networks in autism spectrum conditions," *Social cognitive and affective neuroscience,* 2013.

[339] R. Fergus, Y. Weiss and A. Torralba, "Semi-supervised learning in gigantic image collections.," *In Advances in Neural Information Processing Systems (NIPS).,* 2009.

[340] R. Shams, "Semi-supervised Classification for Natural Language Processing. ," 2014.

[341] Y. Liu and K. Kirchhoff, "Graph-based semi-supervised learning for phone and segment classification.," in *Interspeech,* 2013.

[342] J. S. Turek, T. L. Willke, P. Chen and P. J. Ramadge, "A semi-supervised method for multi-subject FMRI functional alignment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , New Orleans, 2017.

[343] D.-H. Lee, "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks," in *ICML 2013 Workshop : Challenges in Representation Learning (WREPL).*, 2013.

[344] M. Ranzato and M. Szummer, "Semi-supervised learning of compact document representations with deep networks," in *the 25th International Conference on Machine Learning (ICML),* 2008.

[345] J. Weston, F. Ratle, H. Mobahi and R. Collobert, "Deep learning via semi-supervised embedding," *Neural Networks: Tricks of the Trade,* p. 639–655, 2012.

[346] U. Mahmood, M. Rahman, Mahfuzur, A. Fedorov, Z. Fu, V. Calhoun and S. Plis, "Learnt dynamics generalizes across tasks, datasets, and populations," 2019.

[347] 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, no. 10, p. 1016, 2018.

[348] M. Belhaj, P. Protopapas and W. Pan, "Deep Variational Transfer: Transfer Learning through Semi-supervised Deep Generative Models. ," 2018.

[349] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko and T. & Darrell, "Deep Domain Confusion: Maximizing for Domain Invariance. ," 2014.

[350] B. Gholami, P. Sahu, O. Rudovic, K. Bousmalis and V. Pavlovic, *Unsupervised multi-target domain adaptation: An information theoretic approach.*

[351] J. Hoffman, M. Mohri and N. Zhang, "Algorithms and theory for multiple-source adaptation," *in Advances in Neural Information Pro- cessing Systems,* p. 8246–8256, 2018.

[352] S. Zhou, C. Cox and H. Lu, "Improving Whole-Brain Neural Decoding of fMRI with Domain Adaptation," 2019.

[353] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. de Leeuw, C. M. Tempany and B. van Gin- neken, "Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, p. 516–524, 2017.

[354] P. Li, Y. Ying and C. Campbell, "A variational approach to semi-supervised clustering," in *In Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2009.

[355] Y. Wang, G. Haffari, S. Wang and G. aMori, "A rate distortion approach for semi-supervised conditional random fields," in *In Advances in Neural Information Processing Systems (NIPS)*, 2009.

[356] J. Gordon and J. Hernández-Lobato, "Bayesian Semisupervised Learning with Deep Generative Models ," 2017.

[357] M. Tschannen, O. Bachem and M. Lucic, "Recent Advances in Autoencoder-Based Representation Learning ," 2018.

[358] M. Belhaj, P. Protopapas and W. Pan, "Deep Variational Transfer: Transfer Learning through Semi-supervised Deep Generative Models. ," 2018.

[359] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," *In Advances in neural information processing systems*, p. 1313– 1320, 2009.

[360] C. Louizos, K. Swersky, Y. Li, M. Welling and R. Zemel, "The variational fair autoencoder," in *International Conference on Learning Representations (ICLR)*, 2016.

[361] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, p. 2579–2605, 2008.

[362] K. Miller, F. Alfaro-Almagro and N. Bangerter, "Multimodal population brain imaging in the UK Biobank prospective epidemiological study," *Nat. Neurosci*, vol. 19, p. 1523–1536, 2016. .

[363] D. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. Curtiss, S. Della Penna, D. G. M. H. N. H. A. L.-P. L. Feinberg, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, Petersen, S.E., F. Prior, B. Schlaggar, S. Smith, A. Snyder, J. Xu and E. Yacoub, "The Human Connectome Project: a data acquisition perspective.," *Neuroimage*, vol. 62, p. 2222–2231, 2012.

[364] C. RC and J. GA, "A whole brain fMRI atlas generated via spatially constrained spectral clustering," *Hum Brain Mapp*, vol. 33(8), pp. 1914-1928, 2012.

[365] L. Snoek, M. Miesen, T. Beemsterboer, A. Leij and H. & Scholte, "The Amsterdam Open MRI Collection (AOMIC): A Collection of Publicly Available Population Imaging Datasets," 2019.

[366] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, J. W. J. Jr, L. M. Shaw, A. W. Toga, J. Q. Trojanowski and M. W. Weiner, "Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization.," *Neurology*, vol. 74(3), p. 201–209, 2010.

[367] N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard and M. W. Feldman, "Clines, clusters, and the effect of study design on the inference of human population structure," *PLoS genetics,* vol. 1(6), p. e70, 2005.

[368] A. Holmes, M. Hollinshead and T. O'Keefe, "Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures.," *Sci Data 2, 150031 ,* 2015.

[369] S. TD, C. JJ and R. K, "The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth.," *Neuroimage.(Pt B),* pp. 1115-1119, 2016;124.

[370] L. Jingmei, W. Weifei, D. Xue and G. Peng, "Multi-Source Deep Transfer Neural Network Algorithm," *Sensors ,* vol. 19. 3992. , 2019.

[371] O. Esteban, C. Markiewicz and R. Blair, "fMRIPrep: a robust preprocessing pipeline for functional MRI," *Nat Methods,* vol. 16, p. 111–116, 2019.

[372] J. P. Fortin, D. Parker, B. Tunç, T. Watanabe, M. A. Elliott, K. Ruparel, D. R. Roalf, T. D. Satterthwaite, R. C. Gur, R. E. Gur, R. T. Schultz, R. Verma and R. T. Shinohara, "Harmonization of multi-site diffusion tensor imaging data," *NeuroImage,* vol. 161, p. 149–170, 2017.

[373] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, M. McInnis, M. L. Phillips, M. H. Trivedi, M. M. Weissman and R. T. Shinohara, "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage,* vol. 167, pp. 104-120, 2018.

[374] P. S, K. SI and F. E, "Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease.," *Med Image Anal.,* vol. 48, pp. 117-130, 2018.

[375] J. Zhang, C. Zhang, L. Yao, X. Zhao and Z. Long, "Brain State Decoding Based on fMRI Using Semisupervised Sparse Representation Classifications. Computational Intelligence and Neuroscience," *Computational Intelligence and Neuroscience. ,* p. 2018:3956536, 2018.

[376] C. chun, Z. a, Q. jie and Y. Zhaoa, "Transferable attention networks for adversarial domain adaptation," vol. 539, pp. 422-433, 10 2020.

[377] L. Chen, Y. Yang, J. Wang, W. Xu and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation.," *CVPR,* p. 3640–3649, 2016.

[378] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang and X. Tang, "Residual attention network for image classification.," *CVPR,* p. 6450–6458, 2017.

[379] S. Moon and J. G. Carbonell, "Completely heterogeneous transfer learning with attention - what and what not to transfer," *IJCAI,* p. 2508–2514, 2017.

[380] I. Redko, E. Morvant, A. Habrard, M. Sebban and Y. Bennani, "Advances in Domain Adaptation Theory," *ISTE Press - Elsevier,* p. 187, 2019.

[381] C. Horien, S. Noble and A. Greene, " A hitchhiker's guide to working with large, open-source neuroimaging datasets.," *Nat Hum Behav,* 2020.

[382] X. Li, Y. Gu, N. Dvornek, L. Staib, P. Ventola and J. Duncan, "Multi-site fMRI Analysis Using Privacy-preserving Federated Learning and Domain Adaptation: ABIDE Results," *Medical Image Analysis,* p. 101765, 2020.

[383] J. P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, M. McInnis, M. L. Phillips, M. H. Trivedi, M. M. Weissman and R. T. Shinohara, "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage*, vol. 167, p. 104–120, 2018.

[384] D. Pedamonti, "Comparison of non-linear activation functions for deep neural networks on MNIST classification task," *ArXiv*, vol. 1804.02763, 2018.

[385] W. M. Kouw and L. M., "A review of domain adaptation without target labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[386] S. R. Panta, R. Wang, J. Fries, R. Kalyanam, N. Speer, M. Banich, K. Kiehl, M. King, M. Milham, T. D. Wager, J. A. Turner, S. M. Plis and V. D. Calhoun, "A Tool for Interactive Data Visualization: Application to Over 10,000 Brain Imaging and Phantom MRI Data Sets," *Frontiers in neuroinformatics*, vol. 10, no. 9, 2016.

[387] P. Washington, K. M. Paskov, H. Kalantarian, N. Stockham, C. Voss, A. Kline, R. Patnaik, B. Chrisman, M. Varma, Q. Tariq, K. Dunlap, J. Schwartz, N. Haber and D. P. Wall, "Feature Selection and Dimension Reduction of Social Autism Data. Pacific Symposium on Biocomputing.," *Pacific Symposium on Biocomputing*, vol. 25, p. 707–718, 2020.

[388] M. Rakić, M. Cabezas, K. Kushibar, A. Oliver and X. Lladó, "Improving the detection of autism spectrum disorder by combining structural and functional MRI information. ," *NeuroImage : Clinical*, vol. 25, 2020.

[389] T. P. DeRamus, B. S. Black, M. R. Pennick and R. K. Kana, "Enhanced parietal cortex activation during location detection in children with autism. ," *Journal of neurodevelopmental disorders*, vol. 6, no. 1, p. 37, 2014.

[390] R. Newman-Norlund, H. van Schie and A. van Zuijlen, "The mirror neuron system is more active during complementary compared with imitative action.," *Nat Neurosci 10*, p. 817–818, 2007.

[391] M. D'Esposito, J. A. Detre, G. K. Aguirre, M. Stallcup, D. C. Alsop, L. J. Tippet and M. J. Farah, "A functional MRI study of mental image generation.," *Neuropsychologia*, vol. 35, no. 5, p. 725–730, 1997.

[392] J. T. Serences and S. Yantis, "Selective visual attention and perceptual coherence.," *Trends in cognitive sciences*, vol. 10, no. 1, p. 38–45, 2006.

[393] M. Corbetta and G. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nat Rev Neurosci*, vol. 3, no. 3, p. 201–215, 2002.

[394] Q. Gu, Z. Li and J. Han, "Joint feature selection and subspace learning," *IJCAI 2011*, vol. 22, p. 1294, 2011.

[395] J. Li, J. Zhao and K. Lu, "Joint Feature Selection and Structure Preservation for Domain Adaptation," *IJCAI*, 2016.

[396] S. Ben-david, J. Blitzer, K. Crammer and O. Pereira, "Analysis of representations for domain adaptation," *In NIPS*, 2007.

[397] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai and K. Keutzer, "Multi-source domain adaptation for semantic segmentation," *in Advances in Neural Information Processing Systems*, p. 7285–7298, 2019.

[398] D. Durstewitz, G. Koppe and A. Meyer-Lindenberg, "Deep neural networks in psychiatry," *Mol Psychiatry*, vol. 24, p. 1583–1598, 2019.

[399] K. Saenko, B. Kulis, M. Fritz and T. Darrell, "Adapting visual category models to new domains," in *European conference on computer vision*, Berlin, 2021.

[400] S. J. Pan, I. W. Tsang, J. T. Kwok and Q. Yang, "Domain adaptation via transfer component analysis.," *IEEE transactions on neural networks,* vol. 22, no. 2, pp. 199-210, 2010.