Autism Classification Using Topological Features and Deep Learning: A Cautionary Tale*

Archit Rathore[1], Sourabh Palande[1], Jeffrey S. Anderson[1], Brandon A. Zielinski[1], P. Thomas Fletcher[2], and Bei Wang[1]

[1] University of Utah, Salt Lake City, UT 84112, USA
{archit, sourabh, beiwang}@sci.utah.edu, andersonjeffs@gmail.com, brandon.zielinski@utah.edu
[2] University of Virginia, Charlottesville, VA 22904-4259, USA
ptf8v@virginia.edu

Abstract. The identification of autistic individuals using resting state functional connectivity networks can provide an objective diagnostic method for autism spectrum disorder (ASD). The present state-of-the-art machine learning model using deep learning has a classification accuracy of 70.2% on the ABIDE (Autism Brain Imaging Data Exchange) data set. In this paper, we explore the utility of topological features in the classification of ASD versus typically developing control subjects. These topological features have been shown to provide a complementary source of discriminative information in applications such as 2D object classification and social network analysis. We evaluate the performance of three different representations of topological features - persistence diagrams, persistence images, and persistence landscapes - for autism classification using neural networks, support vector machines and random forests. We also propose a hybrid approach of augmenting topological features with functional correlations, which typically outperforms the models that use functional correlations alone. With this approach, even with a simple 3-layer neural network, we are able to achieve a classification accuracy of 69.2% on the ABIDE data set. However, our experiments also show that the improvement due to topological features is not always statistically significant. Therefore, we offer a cautionary tale to the practitioners regarding the limited discriminative power of topological features derived from fMRI data for the classification of autism.

Keywords: Topological data analysis · Autism classification · Neural networks

1 Introduction

Autism Spectrum Disorder (ASD) is a complex developmental disorder characterized by constant challenges in social interactions and communication, often accompanied by restrictive or repetitive behaviors. Many neuroimaging studies (e.g. [15]) relating these behavior deficits to abnormalities in structural and

* This work was supported in part by NSF IIS 1513616 and NIH R01EB022876.

functional connectivity of the brain have provided an impetus to build machine learning models to classify and characterize ASD.

ASD Classification. Autism Brain Imaging Data Exchange (ABIDE) is a large multi-site repository of brain imaging data for autism studies [5,8]. Several different classification models have been applied to the ABIDE data set with reported accuracy between 60% to 70% [1,15]. For instance, Abraham et al. [1] have used support vector and ridge classifiers and reported 67% prediction accuracy on the full ABIDE data. Advances in artificial neural networks (NN) have opened up a new line of research in the study of autism. Various NN models [10,11] have been proposed for classification of autistic subjects. Among all these, the best classification accuracy of 70.2% is achieved by Heinsfeld et al. [11] using the full ABIDE data, where stacked denoising autoencoders are used to learn robust feature representations which are then used to train a multi-layer perceptron classifier.

Topological Data Analysis. In recent years, methods from computational topology have emerged in machine learning that focus on inferring topological features from data. In particular, *persistent homology* [9], a central tool in topological data analysis, allows us to capture the evolution of topological features (such as components, tunnels, voids, etc.) of the data across multiple scales. These topological features are summarized by the *persistence diagram*, a finite multi-set of points in the plane, which yields a complete description of the topological information in the data. Functional connectivity networks derived from resting state fMRIs can be characterized using such persistence diagrams [18]. However, persistence diagrams do not have the structure of an inner product space (i.e. Hilbert space), making it difficult to interface them with machine learning. Several topological kernels have been proposed in the literature [7,13,14,17], making persistence diagrams suitable for kernel support vector machines (SVM). More recently, Hofer et al. [12] have proposed a NN architecture which takes persistence diagrams as input and learns a task-optimal representation of topological features during training.

Overview. In this paper, we explore the use of topological features derived from functional connectivity networks in the classification of autism. Using two ABIDE datasets, we interface three different representations of topological features – persistence diagrams, persistence images [2] and persistence landscapes [4] – with NN, SVM and random forests (RF). In particular, we propose a hybrid approach that combines the correlation matrices with topological features in a unified pipeline for both SVM and NN. Our experiments show that this approach performs better than those using correlations alone. This suggests that topological features hold some discriminative power in the context of autism classification. However, our experiments also indicate that such an improvement is not always statistically significant. Although our experiments are not exhaustive, we hope that this paper will serve as a cautionary reference for others looking to employ topological data analysis in neuroimaging.

2 Technical Background

Persistence diagram. *Persistent homology* [9] measures how the topological features of data evolve across a varying scale parameter α, where connected components, tunnels and voids are considered as 0-, 1- and 2-dimensional features. In a typical setting, we begin with a point cloud $\mathbb{X} \in \mathbb{R}^d$ in a metric space, denoted as $(\mathbb{X}, d_{\mathbb{X}})$. For some $\alpha \geq 0$, replace each point $x \in \mathbb{X}$ with a ball of radius α centered at x under the metric $d_{\mathbb{X}}$. As α increases, the union of balls undergoes topological changes, where topological features appear and disappear. Persistent homology associates a life span (i.e. death time minus birth time), the *persistence*, to these features. Fig. 1 shows an example, where $\mathbb{X}$ is a point cloud in $\mathbb{R}^2$ and $d_{\mathbb{X}}$ is the Euclidean distance metric. As the radius α increases, we keep track of topological changes of the union of balls: e.g., the green component appears at $\alpha = 0$, and it disappears when it is merged with the red component at $\alpha = 2.5$; A tunnel appears at $\alpha = 4.2$ and disappears at $\alpha = 5.6$. The topological information of this process can be computed using simplicial complexes (Fig. 1(b)) and summarized as a finite multi-set of points in the plane, called a *persistence diagram* (PD). Each point of the diagram corresponds to a topological feature, and its coordinates (b, d) specify at which scales the feature appears (birth time b) and disappears (death time d). For example, the pink component (a 0-dimensional feature) is mapped to a pink point $(0, 2.5)$ in the diagram given its birth time $b = 0$ and death time $d = 2.5$; and the tunnel (a 1-dimensional feature) is represented by a purple point $(4.2, 5.6)$ in the diagram, see Fig. 1(c).

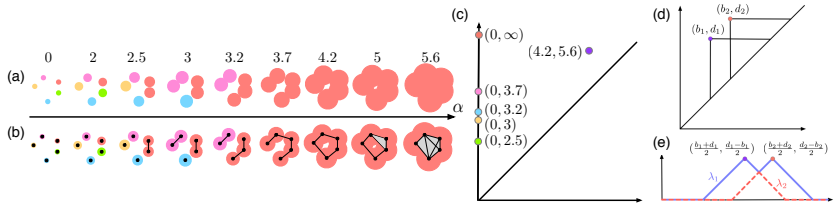

Fig. 1: Left (a)-(c): computing persistent homology of a point cloud in $\mathbb{R}^2$. (a) A nested sequence of topological spaces formed by unions of balls at increasing parameter values. (b) A sequence of simplicial complexes that captures the same topological information as in (a). (c) 0- and 1-dimensional persistence diagrams (the purple point) combined. Right (d)-(e): transform a persistence diagram (d) to a persistence landscape (e).

Persistence Landscape. Persistence landscape (PL) [4] transforms a persistence diagram to a sequence of piece-wise linear functions. Informally, these functions are obtained by first changing the coordinates of points (b, d) to $((b+d)/2, (d-b)/2)$, and then stacking isosceles triangles from each point in the transformed space such that the length of their bases equals the persistence of the point. These piece-wise linear functions can then be sampled uniformly to obtain a discrete vector representation. See Fig. 1 (right) for an example, where functions λ_1 and λ_2 form a PL.

Persistence Images. Persistence images (PI) [2] is another way of transforming persistent diagrams to representations that can be easily vectorized. Informally, they can be thought of as *heatmaps*, generated from a weighted sum of Gaussians centered at each point (b, p) where b is birth and $p = d - b$ is the persistence of a feature in the persistence diagram.

Kernels. Persistence landscapes and persistence images are vector representations that can be used readily in learning algorithms such as SVM under an L_2 metric. To interface persistence diagrams with kernel-based learning, we use kernels defined on persistence diagrams proposed in recent work, namely the persistence scale-space kernel [17], the persistence weighted Gaussian kernel [13], the sliced Wasserstein kernel [7] and the persistence Fisher kernel [14], denoted as K_S, K_G, K_W and K_F respectively. For mathematical formulations, see the supplementary material.

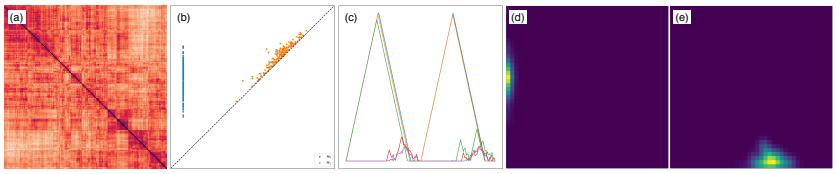

Fig. 2: Different representations of topological features derived from a correlation matrix (a): (b) persistence diagrams; (c) persistence landscapes of 1-dimensional features; (d)-(e) persistence images of 0- and 1-dimensional features.

3 Methods

Data. The ABIDE dataset [5,8] consists of resting state fMRI brain scans from 16 different imaging sites. We use Craddock 200 (**CC200**) data set, where the imaging data is first preprocessed using the Configurable Pipeline for Analysis of Connectomes (CPAC) and then the mean blood oxygenation level dependent (BOLD) signals for 200 regions of interest (ROIs) are extracted. After filtering for missing and invalid data, it contains 1035 subjects, of which 505 are ASD patients and 530 are TDC (typically developing control subjects). We also use Craddock 400 (**CC400**) generated using the same procedure as **CC200** but has 400 ROIs instead.

For each subject, we compute the Pearson correlation between the BOLD signals for all pairs of ROIs, resulting in an $n \times n$ correlation matrix M, where $n = 200$ for **CC200** and $n = 400$ for **CC400**. We then map M to a point cloud $\mathbb{X}$ in a metric space: each ROI is mapped to a point; and the distance between two points is defined as $d_{\mathbb{X}}(x, y) = \sqrt{1 - M(x, y)}$. We compute PD_0 and PD_1, persistence diagrams in 0 and 1 dimension, respectively.

Classification with correlation features. A *correlation feature* vector is obtained by flattening the lower triangular part (below the diagonal) of the cor-

relation matrix, resulting in a vector of length $n(n-1)/2$. Correlation features are naturally equipped with a Euclidean (L_2 distance) metric.

We use three models to perform classification with correlation features alone. We use **scikit-learn** [16] for SVM and RF implementations. For both models, parameter estimation is done using grid search. For SVM ($\mathbf{SVM}_{Corr}$): we use a linear SVM (a linear kernel in our setting is shown to outperform other kernels); a tunable parameter controls the trade-off between misclassification and margin size. For RF ($\mathbf{RF}_{Corr}$): we use 500 trees, each of maximum depth of 5; tunable parameters include the number of trees and the tree depth. For NN ($\mathbf{NN}_{Corr}$): we train a fully connected NN with 3, 5 and 7 layers with ReLU activation. We apply a $p = 0.5$ probability dropout and batch normalization for regularization. The network is trained by optimizing the cross entropy loss using Stochastic Gradient Descent (SGD) with Nesterov momentum. For all NN, we chose a learning rate $r = 0.1$ and momentum $\tau = 0.9$ and train for 200 epochs.

Classification with topological features. Encoding topological information alone can offer alternative perspectives on classification. Since persistence images and persistence landscapes can be easily vectorized, they interface with SVM, RF and NN in a straightforward way. We use Python **Persim** package for persistence images; we use images of dimension 200×200 and 0 spread. We use **sklearn_tda** [6] Python package to compute persistence landscapes of order 1 through 5 sampled at 2000 discrete points. For persistence diagrams, we use persistence scale-space kernel [17] K_S to make them suitable for kernel SVM ($\mathbf{SVM}_{PD}$). To interface persistence diagram with NN ($\mathbf{NN}_{PD}$), we use the approach of Hofer et al. [12]. The key idea is to construct a projection layer (defined independently for PD_0 and PD_1) on top of a NN architecture that takes as input a persistence diagram, defines a projection with respect to a collection of structure elements encoding persistence, and outputs an n-dimensional vector. For more details, see the supplementary material.

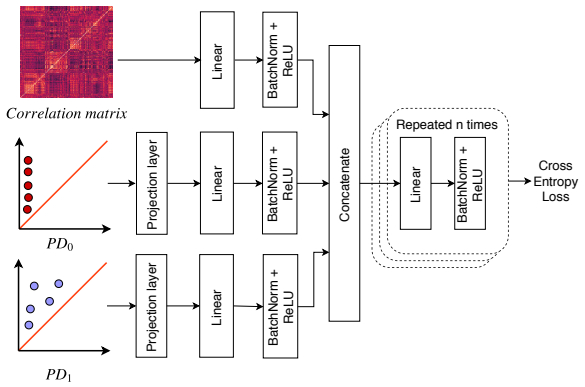

Fig. 3: A NN architecture that combines correlation and topological features.

Classification with combined features. Finally, we propose and train hybrid classifiers using both correlation and topological features. For SVM, we combine both features using a linear combination of kernels [18]. In terms of PD, for instance, the combined kernels are given by: $K_{PD+Corr} = w_0 K_{PD_0} + w_1 K_{PD_1} + (1 - w_0 - w_1) K_{Corr}$ where $w_0, w_1 \geq 0$ and $w_0 + w_1 \leq 1$. The tunable weight parameters are estimated using grid search.

K_{PD_0} and K_{PD_1} are kernels on PD_0 and PD_1; we experiment with four different kernels mentioned in Sec. 2. K_{Corr} denotes the kernel on correlation features. For correlation features, persistence images and persistence landscapes, a kernel is simply the Euclidean dot product between the feature vectors. We also propose a hybrid NN architecture ($\mathbf{NN}_{PD+Corr}$) as shown in Fig. 3 to leverage both correlation and topological features. As described in Sec. 2 and [12], two independent projection layers are defined for PD_0 and PD_1 respectively, each is independently projected onto an n-dimensional vector. Together with a correlation feature vector, these three vectors are independently processed through several fully connected layers of NN in parallel and the outputs are then concatenated to form a single long vector. The final layer is a softmax which takes this concatenated vector as input and generates binary class labels (ASD vs TDC). We train this network by minimizing the cross entropy loss over the training set. The number of intermediate fully connected layers in the network varies and we report results for 3, 5 and 7 layers in Sec. 4.

4 Results

We experiment with various classification models using correlation features, topological features, and their combinations. We summarize performance measurements in Fig. 4 for **CC200** and **CC400**. All results are obtained by running a 5-fold stratified cross-validation scheme where we report the mean classification accuracy over all folds. For abbreviations: *Corr* means using correlation features; *PD*, *PI* and *PL* denote persistence diagrams, persistence images and persistence landscapes, where kernel SVM interfaces with PD via persistence scale space kernel; $\mathbf{NN5}_{PD+Corr}$ means NN 5-layer hybrid model combining correlation with topological features from PD, see the supplementary material for implementation details and additional results. Our main experimental results are as follows. First, our proposed NN 3-layer hybrid models that combine correlation with topological features have the highest mean classification accuracy (Fig. 4, bold and red entries). Adding more layers leads to a decrease in classification accuracy - this may be due to the fact that more layers lead to a larger number of learnable parameters which are hard to train with a relatively small training data. Second, NN with correlation features and NN hybrid models with combined features provide a significant improvement in test accuracy over SVM and RF (Fig. 5). Third, three representations of topological features (PD, PI and PL) have similar performance (Fig. 4); same is true for kernel SVM models using four different kernels (Fig. 5 right). Finally, the improvement due to topological

Model	CC200	CC400	Model	CC200	CC400	Model	CC200	CC400
SVM_{Corr}	65.41	66.33	-	-	-	-	-	-
RF_{Corr}	64.81	63.92	-	-	-	-	-	-
$NN3_{Corr}$	68.35	63.92	-	-	-	-	-	-
$NN5_{Corr}$	68.46	65.58	-	-	-	-	-	-
$NN7_{Corr}$	67.10	62.06	-	-	-	-	-	-
SVM_{PD}	53.03	53.69	SVM_{PI}	54.54	53.76	SVM_{PL}	53.03	53.69
RF_{PD}	-	-	RF_{PI}	52.25	53.04	RF_{PL}	52.51	53.12
$NN3_{PD}$	56.06	55.90	$NN3_{PI}$	58.56	56.10	$NN3_{PL}$	55.36	54.24
$NN5_{PD}$	56.15	56.04	$NN5_{PI}$	59.09	57.39	$NN5_{PL}$	55.18	55.72
$NN7_{PD}$	55.48	54.33	$NN7_{PI}$	56.75	55.58	$NN7_{PL}$	54.85	53.67
$SVM_{PD+Corr}$	65.86	63.36	$SVM_{PI+Corr}$	64.25	62.68	$SVM_{PL+Corr}$	65.65	64.12
$\mathbf{NN3}_{PD+Corr}$	**69.19**	**67.84**	$\mathbf{NN3}_{PI+Corr}$	**67.2**	**67.02**	$\mathbf{NN3}_{PL+Corr}$	**68.5**	**66.76**
$NN5_{PD+Corr}$	68.20	66.03	$NN5_{PI+Corr}$	66.87	66.23	$NN5_{PL+Corr}$	67.45	66.48
$NN7_{PD+Corr}$	64.47	61.25	$NN7_{PI+Corr}$	65.10	64.16	$NN7_{PL+Corr}$	67.02	65.26

Fig. 4: Mean classification accuracy using various classifiers and feature combinations.

	RF_{Corr}	SVM_{Corr}	$SVM_{PD+Corr}$	$NN3_{Corr}$
SVM_{Corr}	0.1502			
$SVM_{PD+Corr}$	0.1943	0.4213		
$NN3_{Corr}$	**0.0461**	**0.0480**	0.0631	
$NN3_{PD+Corr}$	**0.0406**	**0.0446**	**0.0414**	0.1894
	RF_{Corr}	SVM_{Corr}	$SVM_{PI+Corr}$	$NN3_{Corr}$
$SVM_{PI+Corr}$	0.1943	0.4213		
$NN3_{Corr}$	-	-	**0.0420**	
$NN3_{PI+Corr}$	**0.0493**	0.0763	0.0734	0.7432
	RF_{Corr}	SVM_{Corr}	$SVM_{PL+Corr}$	$NN3_{Corr}$
$SVM_{PL+Corr}$	0.1623	0.3513		
$NN3_{Corr}$	-	-	0.0581	
$NN3_{PL+Corr}$	**0.0467**	0.0683	0.0717	0.3524

Kernel SVM	CC-200	CC-400
K_S	53.03	53.69
K_S + Corr	65.86	63.36
K_G	52.51	53.12
K_G+ Corr	62.98	61.41
K_W	55.36	54.24
K_W + Corr	64.73	64.12
K_F	55.18	55.72
K_F + Corr	61.48	60.25

Fig. 5: Left: the statistical significance of improvements in classification accuracy, comparing each row method against each column method, captured by p-values. Right: mean accuracy for kernel SVM with different topological kernels.

features is not always statistically significant (Fig. 5 left), especially comparing NN hybrid models against NN with correlation features.

We use permutation tests to determine the statistical significance of the improvement shown in Fig. 4. The test statistic we use is the difference in accuracy of predictions from two models ($\text{Accuracy}_{model_1}$ - $\text{Accuracy}_{model_2}$). Since the number of labels is too large to compute all possible permutations, we estimate the p-values using a Monte-Carlo approximation with $100,000$ samples. In each permutation, we perform random pairwise swaps of labels predicted by $model_1$ and $model_2$ for the subjects and compute the new test statistic. The p-value is the fraction of samples where the test statistic of the permuted labels is greater than the test statistic for the original labels. For fully connected NN and hybrid NN models, adding layers reduces accuracy of classification, hence we report the p-values for 3 layer models only. The results are shown in Fig. 5 (left), which contains p-values that capture statistical significance of improvement in accuracy, by comparing each row method with each column method.

5 Discussion

Our goal is to explore the utility of topological features in the classification of autism. We observe modest improvement in classification accuracy using hybrid models that combine topological and correlation features; it shows that there is information not conveyed in correlations themselves. However, our results combined with the ones reported by others [1,11,15] lead us to conjecture that 70% might be the best classification accuracy any model can achieve on the *entire* ABIDE data set. Noise in the data and heterogeneity across multiple sites are only part of the reason behind the relatively low classification accuracy. ABIDE data consists of rs-fMRI scans acquired with low temporal resolution and short acquisitions sequences ($\leq$ 10 min/subject) with relatively poor single subject reliability. It is possible that the added advantage of topological features could become much more important if the image acquisition strategy included lower-noise, longer-duration acquisitions. There is a very wide developmental age range in the ABIDE data, which also contributes to the challenge. Perhaps age is not a linear factor in considering topological features, and performance might be better in samples with narrower age range. Finally, autism itself is a markedly heterogeneous disorder. Grouping patients under a single label may not be as powerful as one that can discriminate subsets of patients. So high diagnostic accuracy may not be as important as subphenotyping different sets of patients who may have different treatment needs. Topological features may provide a deeper insight into subtyping in autism.

References

1. Abraham, A., et al.: Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. NeuroImage **147**, 736–745 (2017)
2. Adams, H., et al. : Persistence images: A stable vector representation of persistent homology. Journal of Machine Learning Research **18**(1), 218–252 (2017)
3. Bubenik, P.: Statistical topological data analysis using persistence landscapes. Journal of Machine Learning Research **16**(1), 77–102 (2015)
4. Cameron, C., et al.: The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Frontiers in Neuroinformatics **7** (2013)
5. Carriére, M.: sklearn-tda: a scikit-learn compatible python package for machine learning and tda. https://github.com/MathieuCarriere/sklearn-tda
6. Carrière, M., Cuturi, M., Oudot, S.: Sliced Wasserstein kernel for persistence diagrams. Proceedings of the 34th International Conference on Machine Learning **70**, 664–673 (2017)
7. Di Martino at al.: The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry **19**(6), 659–667 (2014)
8. Edelsbrunner, H., Harer, J.: Persistent homology–a survey. Contemporary mathematics **453**, 257–282 (2008)
9. Guo, X., Dominick, K.C., Minai, A.A., Li, H., Erickson, C.A., Lu, L.J.: 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 **11**, 460 (2017)
10. Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., Meneguzzi, F.: Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage: Clinical **17**, 16–23 (2018)
11. Hofer, C., Kwitt, R., Niethammer, M., Uhl, A.: Deep learning with topological signatures. Advances in Neural Information Processing Systems pp. 1634–1644 (2017)
12. Kusano, G., Fukumizu, K., Hiraoka, Y.: Kernel method for persistence diagrams via kernel embedding and weight factor. Journal of Machine Learning Research **18**(1), 6947–6987 (2017)
13. Le, T., Yamada, M.: Persistence fisher kernel: A Riemannian manifold kernel for persistence diagrams. Advances in Neural Information Processing Systems **31**, 10028–10039 (2018)
14. Nielsen, J. A., et al.: Multisite functional connectivity MRI classification of autism: ABIDE results. Frontiers in Human Neuroscience **7**, 599 (2013)
15. Pedregosa et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
16. Reininghaus, J., Huber, S., Bauer, U., Kwitt, R.: A stable multi-scale kernel for topological machine learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 4741–4748 (2015)
17. Wong, E., Palande, S., Wang, B., Zielinski, B., Anderson, J., Fletcher, P.T.: Kernel partial least squares regression for relating functional brain network topology to clinical measures of behavior. IEEE International Symposium on Biomedical Imaging pp. 1303–1306 (2016)