Identifying Autism Spectrum Disorder Using Brain Networks: Challenges and Insights

Keanelek Enns
Computer Science
University of Victoria
Victoria, Canada
keanelekenns@uvic.ca

Venkatesh Srinivasan
Computer Science
University of Victoria
Victoria, Canada
srinivas@uvic.ca

Alex Thomo
Computer Science
University of Victoria
Victoria, Canada
thomo@uvic.ca

***Abstract*—Autism Spectrum Disorder (ASD) affects a large portion of the global population both directly and indirectly. The biological etiology of the disorder is not sufficiently understood, and current diagnoses rely on behavioural indicators which do not provide a reliable basis for diagnosis until about 2 years of age. Identifying a biological marker of ASD would aid in understanding the disorder and potentially allow for earlier, more objective diagnoses and treatments to improve the quality of life of individuals possessing ASD. The analysis of functional connectivity in the brain using functional Magnetic Resonance Imaging (fMRI) has been identified as a promising method for discovering such biological markers.**

This study recreated a prominent state-of-the-art work in explainable classification of brain networks, but found results inconsistent with what was claimed. The methods were modified in various ways to improve accuracy and performance. A new, simpler method named Discriminative Edges (DE) was developed which achieved similar accuracies with improved performance and explainability. DE was also adapted to receive raw correlation matrices as well as thresholded correlation matrices representing brain networks, and it was found that raw correlation matrices provided more useful information for classification. An implementation package was provided to aid future researchers in validating and improving upon these results. Suggestions for future work based on the findings of this study were provided, the most important being to procure more datasets, discover data-driven subcategories of ASD, and maintain reproducibility in studies.

***Index Terms*—neuroimaging, fMRI, machine learning, diagnosis, brain networks**

I. Introduction

Reliably diagnosing and understanding Autism Spectrum Disorder (ASD) from a biological perspective poses a difficult challenge due to the disorder's complexity and varied forms of expression [1]–[4]. The criteria for diagnosing ASD, as given by the DSM-5, are vague compared to other neurological disorders [5] and even underwent significant changes in 2013 [6]. Moreover, the disorder is common and affects a large portion of the population [7].

ASD is currently diagnosed by observing behaviour [5] and cannot reliably be diagnosed until an individual reaches about 2 years of age [8]. An early diagnosis can be crucial to getting proper support for an individual and providing caretakers with an understanding of the condition that will improve the individual's quality of life [9]. However, in many cases, a diagnosis is not given until much later.

If features of the brain during early development could reliably identify an individual with ASD, it would not only provide the early diagnoses sought after, but it would help neuroscientists gain a better understanding of what causes ASD biologically, which could lead to a plethora of methods for improving the quality of life and healthy development of such individuals.

Unfortunately, such features of the brain have been elusive to researchers despite the surge of effort in this area in recent years [10]–[13]. Thus, the search continues in order to address this important issue and gain a better understanding of the pervasive, yet misunderstood disorder.

The focus of this study is to work towards a better method of identifying and diagnosing ASD without relying on behavioural information, but rather by using brain imaging data.

A common approach for deriving useful information from brain scans, such as those produced by fMRI, is to divide the brain into regions of interest (ROIs) based on their functionality and construct a graph whose nodes correspond to ROIs and whose edges correspond to correlations of brain activity between ROIs. This turns the problem of classifying fMRI scans into a graph classification problem.

Machine learning (ML) and artificial intelligence (AI) have been shown to out-perform humans significantly in a multitude of domains [14]–[17], and the domain of graph classification is no exception [18]. But how is performance measured for graph classification? Metrics such as accuracy, precision, and recall are essential for evaluating any classifier [19], and there is no doubt that ML and AI models can achieve impressively high scores in such areas. However, recently there has been a trend towards *explainability* in the AI world [20], [21].

This is because industries, governments, and organizations, especially those that deal with critical decision making such as the medical field, are hesitant to adopt prediction models without knowing *how and why* they make decisions, regardless of how accurate these models are reported to be [22]. Moreover, emphasizing explainability can provide insights that may not have been detected through classical methods and may lead to further advancements in research. Therefore, it is increasingly important to find classification models that are explainable and simple to understand, while also achieving high accuracy, precision, and recall scores.

In their paper, "Explainable Classification of Brain Networks via Contrast Subgraphs", Lanciano *et al.* proposed a method for translating the previously described brain networks into two-dimensional vectors with a simple interpretation [23]. This translation of a graph into a simpler representation is known as a graph embedding [24]. The graph embedding employed by Lanciano *et al.* involves thresholding correlation values of constructed brain networks and the use of contrast subgraphs (CSs).

This study sought to assess the current state of research in this area, improve upon existing methods, and provide insights regarding possible directions of future work. The following contributions were made:

- A replication of the novel CS method by Lanciano *et al.* along with various modifications to the CS method.
- A new approach to the problem named Discriminative Edges (DE) which provides a simpler solution with a fraction of the running time of the CS method.
- An implementation of an effect size thresholding approach for comparison [25] with our DE method.
- An implementation package to recreate all the work done in this study[1] with suggestions for the future of this area of research.

II. Related Works

The problem of classifying ASD using fMRI data has been studied intensively over the past decade [10]–[13], [26], [27]. Constructing correlation matrices (or brain networks) from blood-oxygen-level-dependent (BOLD) signals in fMRI images, as seen in Figure 1, is fairly common in this area of research [18], [28]–[31]. This is because the raw four-dimensional fMRI data is too large to be useful for most ML strategies. Thomas *et al.* attempted to reduce the size of the data by collapsing the temporal dimension using various metrics such as regional homogeneity. Their accuracies on the full ABIDE I datasets reached about 66% [32].

Subah *et al.* report a high accuracy of 88% on the ABIDE I dataset using a Dense Neural Network (DNN) and the BASC brain atlas [29], however, they do not focus on explainability or utilize feature engineering which can be useful for understanding as well as performance [30]. Many other studies using deep learning (DL) models similarly focus on fine-tuning model architectures rather than feature engineering [31] or they use complex feature selection techniques that are not interpretable to individuals that are not machine learning experts [27], [30].

On the other hand, Kong *et al.* use a measure called F-score to select the top 3000 correlation values as features. They use a DNN and a small subset of the ABIDE I dataset and claim an accuracy of 90% [18]. Similarly, Iidaka uses effect-size thresholding to select features before employing a Probabilistic Neural Network (PNN) for classification. They also use a subset of the ABIDE I dataset (individuals under 20 years of age) and claim about 90% accuracy [25].

[1]https://github.com/keanelekenns/brain-network-classification

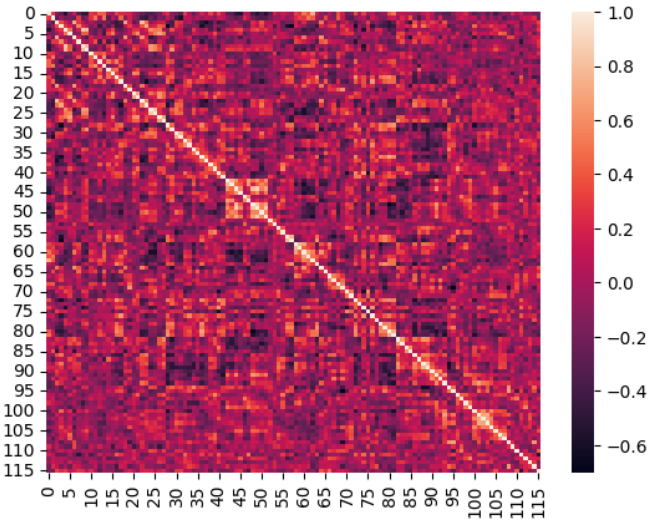


Fig. 1: This matrix contains the pairwise Pearson correlation coefficients of every ROI in the brain of an individual from the ABIDE I dataset using the AAL brain atlas [1]. The matrix is symmetric, and the main diagonal contains zeros as the correlation of each ROI to itself is irrelevant. This can be viewed as a weighted, undirected brain network where edges represent the strength of functional connections in the brain.

It has been found that studies, such as the previously mentioned studies, done on small subsets of data report higher accuracies than those that use larger datasets, and more specifically, the difference has been noted between single-site and multi-site studies, possibly due to varying experimental conditions and extraction methods [33]. It is also likely that over-fitting occurs in studies with small datasets; such models are not extensible to new datasets.

Many approaches rely on a single brain atlas to abstract raw fMRI data. However, there is no single brain atlas that is considered superior. In a recent study, Epalle *et al.* successfully utilized information from multiple brain atlases simultaneously to generate predictions on the ABIDE I dataset [34]. For each atlas, they followed a similar approach to others with respect to generating correlation matrices from the BOLD time series of ROIs. They then selected a fixed-size set of edges in the correlation matrices derived from each atlas and fed them into a multi-input single-output deep neural network. Their experiments showed an improvement in performance over similar deep-learning pipelines using fewer atlases.

There has been a recent trend towards more explainable AI (sometimes referred to as XAI) [35] and some tools are available for explaining the predictions of black-box models [20], [21], but they have various limitations, primarily in the form of computational complexity. Perotti *et al.* created a tool for deriving SHAP values in the domain of graph classification by using motifs as features [36], but determining these explanations is computationally expensive. Similarly, Abrate and Bonchi employed a strategy to find counterfactual

graphs to help explain black-box classifiers, but the process is computationally expensive and only reflects what the classifier deems as important information whether that information is truly useful for classification or not [37].

The present study was heavily influenced by the work of Lanciano *et al.* as it began with a replication of their paper [23]. They claim an accuracy of 86% on a subset of the ABIDE I dataset comprised of children. They focus on explainability and simplicity of features to assist neuroscientists in interpreting the findings rather than creating a highly accurate classifier that is difficult to understand. The emphasis on explainability is maintained in this study, because, while early diagnosis is the primary goal of this research, neuroscientists and field experts need to be able to interpret the predictions of automated classifiers before they can trust them and learn from them.

III. Contrast Subgraphs

In 2020, Lanciano *et al.* sought to find subgraphs in brain networks called *Contrast Subgraphs* (CSs) that could be used to discriminate between the two classes (i.e. ASD and the typically developed control group labelled TD). A CS is defined as a subset of vertices that induces a dense subgraph in one graph and a sparse subgraph in another, assuming that the graphs share a common vertex set (which is the case for ROIs of brain networks defined on a common brain atlas).

The first step of the approach is to reduce the noise in the input data by applying a threshold to each correlation matrix which causes the brain networks to become unweighted compared to the raw correlation matrices.

In order to derive class specific information, summary graphs are defined for each class over the common set of ROIs representing nodes of the graphs. These are the graphs from which the CSs are discovered. The edge weights in the summary graphs represent the percentage of brain networks in the corresponding class that possess the corresponding edge. For example, if half of the brain networks in the ASD class contain edge e, then edge e will have a value of 0.5 in the ASD summary graph.

In the context of Lanciano *et al.*'s work, finding a contrast subraph between the two summary graphs is equivalent to finding a dense subgraph in their difference. Therefore, a difference network is obtained by subtracting one summary graph from another. Two such difference networks can be created by reordering the subtraction.

Finding dense sugbraphs is a common problem in graph theory, and as such, Lanciano *et al.* repurposed the work of Cadena *et al.* to identify dense subgraphs in the difference networks using a semi-definite programming (SDP) solver [38]. The solver approximates the optimal solution based on their definition of density and a local search algorithm adapted from the work of Tsourakakis *et al.* is used to further refine the solution [39]. The result is a set of nodes that comprises a dense subgraph in a difference network, which represents a CS in this context.

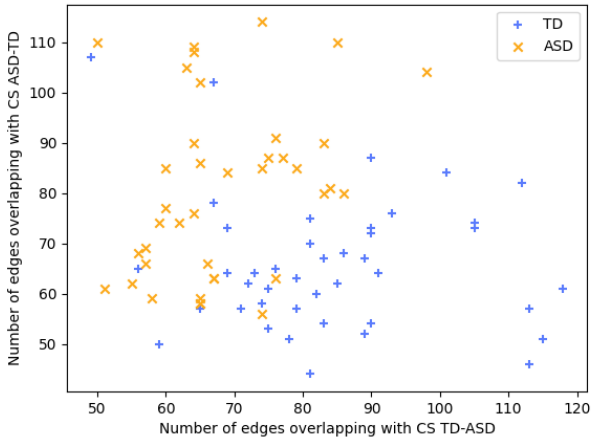


Fig. 2: A group of brain networks plotted in two dimensions based on their features derived from the CSP1 approach.

A CS is found for both of the difference networks, hence each CS represents a group of ROIs that were found to be more connected in one class than the other. Finally, each brain network is translated into a two-feature vector by counting the overlap between each CS as shown in Figure 2. The brain networks used for training are translated first and used to train a classifier, and the remaining brain networks are translated after and predicted using the classifier.

Lanciano *et al.* also define a symmetric variant of the problem. We refer to the already described approach as CSP1, and the variant CSP2. In CSP2, a single CS is found in the difference network containing absolute valued edge weights. The CS is used to induce subgraphs in both of the summary graphs as well as each individual brain network. The distances, computed as the L_1 norms, from the induced brain networks to each induced summary graph are then used as the two features for this approach.

For further details about the CSP1 and CSP2 approaches, we refer the reader to Lanciano *et al.*'s paper. During the replication of their work, however, some modifications were made to their approaches to achieve better computational performance. The modifications include using a quadratic programming (QP) solver rather than an SDP solver when approximating the densest subgraph, improving the logical implementation of the local search algorithm used to refine the solution, and finding multiple contrast subgraphs for each class rather than one. The first two improvements serve to decrease the computation time, while the third improvement utilizes more information from the correlation matrices.

IV. Discriminative Edges

Explainability is the major advantage of the contrast subgraph approach. There are only two features, which makes it easy to visualize the representation of each brain network. In turn, the decisions made by a classifier can be understood by humans.

A disadvantage is made apparent when considering the usefulness of CSs in light of how difficult they are to find. In Section 5.1 of their paper, Lanciano *et al.* showed that the weighted degrees of the nodes in each of the classes' summary graphs were nearly identical [23]. This indicated that there was no clear difference in network structure when looking at the connectivity of certain nodes. The important information comes from the strength of the connections (i.e. the weight of the edges) in the summary graphs. However, CSs are defined as sets of nodes, meaning there could be unimportant edges included within the CSs when inducing a subgraph with them, and those edges are given equal importance in the calculation of the features used for discriminating between the classes.

A simple approach was developed in this study to address this issue. The approach is named *Discriminative Edges* (DE) because it uses the most important edges, or connections in the brain, for discriminating between the two classes. It has the following key features:

- It focuses on connections that discriminate between the classes rather than ROIs, so as not to include unimportant connections in the decision making.
- It has only a linear time complexity for identifying the important edges, and is therefore much faster than the contrast CS approach as well as many others.
- The calculations are simple to understand and trace: it uses dot products and Euclidean vector distances.
- It appropriately weights connections based on their importance.
- It can perform classification on both thresholded or raw correlation matrices.

In the DE approach, a difference network is obtained from two summary graphs, just as in the CS approach. However, rather than approximating the solution to a complex optimization problem by choosing a set of nodes that maximizes an objective function, it simply selects the n most positively weighted edges and n most negatively weighted edges in the difference network (where n is a chosen hyperparameter). Performing this partitioning operation is only linear in time complexity. Note that the magnitudes of the values in the difference network are used to measure the importance or discriminative power of the edges.

In the case of thresholded brain networks, like the ones derived by Lanciano *et al.*, the calculation is very simple. In order to utilize the information of absent edges, each brain network is transformed to have a value of -1 where there is no connection, and 1 where there is a connection. A dot product is then taken between the n positive edges in the individual brain network and the summary graph, and similarly between the n negative edges.

The approach requires slightly different calculations in the case of using raw correlation matrices or weighted brain networks. The summary graphs are calculated by averaging the correlation values of each class's brain networks. Because the networks contain correlation values rather than simple indicators of each connection's presence, it is important to measure similarity to the class's average correlation values rather than whether the values are more positive or negative as in the unweighted case.

Therefore the graphs are treated like vectors and relative similarities are measured with euclidean distances. Let A be the summary graph of the positive class (i.e. the first class in the subtraction to obtain the difference network) and let B be the summary graph of the negative class. Also, let the subscript K denote a set of k edges that induces a subgraph in a brain network or summary graph. Then the relative similarity of an individual brain network i to the positive class is calculated as:

$$\frac{||B_K - i_K|| - ||A_K - i_K||}{||B_K - i_K|| + ||A_K - i_K||} \times 100\% \tag{1}$$

In order to account for the relative importance of each edge, the vectors are scaled using the corresponding difference network values. Finally, the similarity of a brain network with the positive class is calculated by setting K to be the n most positive edges and n most negative edges in the difference network.

Though the two features used for DE provided useful information for classifying brain networks, some information was not being utilized, namely all of the edges between the top and bottom n most discriminative edges. Therefore a third feature was derived for each case: the whole network similarity.

This feature is calculated in a nearly identical way as the first two features in both the unweighted and weighted cases, but it uses every edge to measure an individual brain network's similarity to the positive class. This has the additional benefit of allowing the $2n$ discriminative edges to be compared by their relative importance.

V. Experiments

The experiments in this study were conducted on datasets derived from the ABIDE I initiative. The datasets were provided by the Preprocessed Connectomes Project (PCP), which performed preprocessing of the ABIDE I datasets using a variaty of pipeline tools and parameters. The DPARSF[2] pipeline was used with band-pass filtering and global signal regression, and the version of the AAL[3] brain atlas provided by the PCP was chosen. A description of the categorization and counts of subjects can be found in Table I.

To ensure all approaches were evaluated fairly, a common framework was developed to run the experiments. This framework was implemented similarly to the pipeline module provided by scikit-learn.

The steps of the pipeline consist of a series of transformer classes, each possessing `fit` and `transform` functions, followed by a classifier class, possessing `fit` and `predict` functions.

To limit the variability between approaches, it was decided that their pipelines would only vary in their first step, which

[2]http://preprocessed-connectomes-project.org/abide/dparsf.html

[3]http://preprocessed-connectomes-project.org/abide/Pipelines.html#regions_of_interest

TABLE I: Subject counts by category, file type, and class. The first four categories are defined exactly as in the work of Lanciano *et al.* [23]. The "other" category includes subjects that were not included in any of the first four categories. The "all" category includes all unique subjects. The "Lanciano (thresholded)" column corresponds to the thresholded brain networks provided by Lanciano *et al.* in their repository. The "Raw Correlation" column corresponds to the correlation matrices derived from the PCP-provided BOLD time series.

Category	Lanciano (thresholded)		Raw Correlation	
	ASD	TD	ASD	TD
children	49	52	40	39
adolescents	116	121	114	122
eyesclosed	136	158	141	165
male	420	418	443	455
other	0	0	27	49
all (unique)	457	462	504	551

would receive the brain networks as input and output the feature vector specific to each approach. The second step of the pipeline was the StandardScaler class, which standardizes features by removing the mean and scaling them to unit variance in each dimension. This serves to give each feature of the outputted feature vector a more equal importance in the classification (especially when used in conjunction with a classifier that uses spatial algorithms). Finally, the classifier used for each approach was the SVC[4] model from scikit-learn's SVM module using an RBF kernel.

An additional module was made for performing grid searches for tuning hyperparameters in conjunction with cross-validation and nested cross-validation. In their paper, Lanciano *et al.* describe a nested cross-validation approach with hyper-parameter optimization [23]. It was not clear how the hyperparameters were selected, both with respect to the mechanism for choosing hyperparameters and for evaluating what makes them the "best". Therefore a standard grid search was conducted over the parameters for both the transformer class of the specific approach and the SVC class, which takes two primary hyperparameters: C and gamma. Furthermore, the maximum average accuracy achieved by a given set of hyperparameters over the 5 inner folds of cross-validation on the training data was used to determine the best set of hyperparameters.

A variety of metrics and useful information are outputted for each outer fold of the experiments. These include the following:

- The parameter grid used,
- The chosen parameters,
- The confusion matrix resulting from the predictions,
- Basic metrics such as accuracy, precision, and recall, and
- Average runtimes for various stages of the experiments.

Additionally, for the methods that can be plotted in two or three dimensions (namely the CS methods and DE), three plots are generated from each of the outer folds with the following information:

1) Training points according to their class labels.
2) Test points according to their class labels.
3) Test points according to the predictions that are made.

[4]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

VI. Results

Much effort was put into reproducing the results of Lanciano *et al.*'s work, however, the results came short of those claimed in their paper. Table II shows the results of running 5-fold cross-validation using the approaches described by their work and the best hyperparameters reported in Appendix A of their paper.

TABLE II: Replication results. This is modelled after Table 2 in Lanciano *et al.*'s paper [23] and reports average accuracies with their relative standard deviation in percentages.

	Children	Adolescents	EyesClosed	Male
CSP1	73.5 ± 13.5	60.8 ± 15.6	58.5 ± 9.0	59.3 ± 4.3
CSP2	65.6 ± 14.7	63.7 ± 9.5	58.5 ± 7.6	61.9 ± 4.9

The goal of this study was not exclusively to replicate Lanciano *et al.*'s work. The approaches that were experimented with include the following:

- CSP1-SDP-N1 - The recreation of CSP1 from Lanciano *et al.*'s work. SDP specifies the solver used, and N1 specifies the number of contrast subgraphs used for each feature.
- CSP2-SDP-N1 - The recreation of CSP2 from Lanciano *et al.*'s work.
- CSP1-QP-N3 - The CSP1 approach with modifications. QP specifies the solver used, and N3 specifies the number of contrast subgraphs used for each feature.
- CSP2-QP-N3 - The CSP2 approach with modifications.
- DE - The Discriminative Edges approach.
- Iidaka - A partial recreation of the effect-size thresholding approach of Iidaka [25].

The approaches for CSP1 (both the original and modified versions) can only receive thresholded, unweighted brain networks as input, as it was unclear how to extend the technique to the weighted scenario without changing it significantly. DE and the approaches for CSP2 (both the original and modified versions) can receive both unweighted brain networks and raw correlation matrices. The effect size thresholding approach can only receive raw correlation matrices as input. Hence, there were two kinds of experiments conducted: those receiving unweighted brain networks as inputs (provided by Lanciano *et al.*) and those receiving raw correlation matrices (generated from downloaded BOLD time series in this study).

Numerous empirical experiments were conducted during this study. However, due to the excessive computation time required to evaluate the recreation of Lanciano *et al.*'s methods (which were empirically found to be about four times faster than the original implementations), it was not possible to compare the results of more rigorous experiments such as leave-one-out cross-validation for all methods. The time-consuming approaches were manually tested over smaller parameter grids until an appropriately sized grid could be found that took a reasonable amount of time to run, but gave the approaches

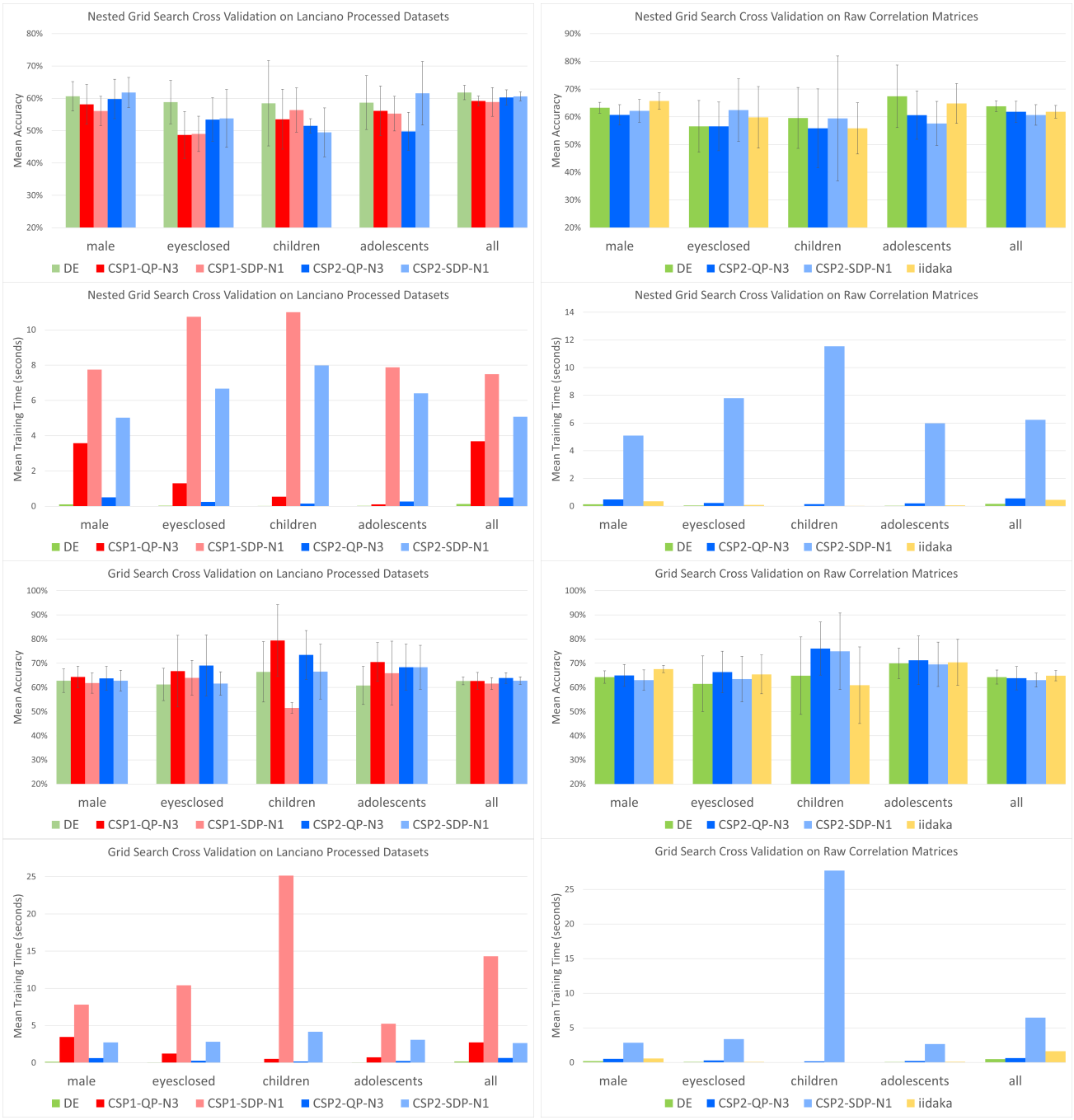
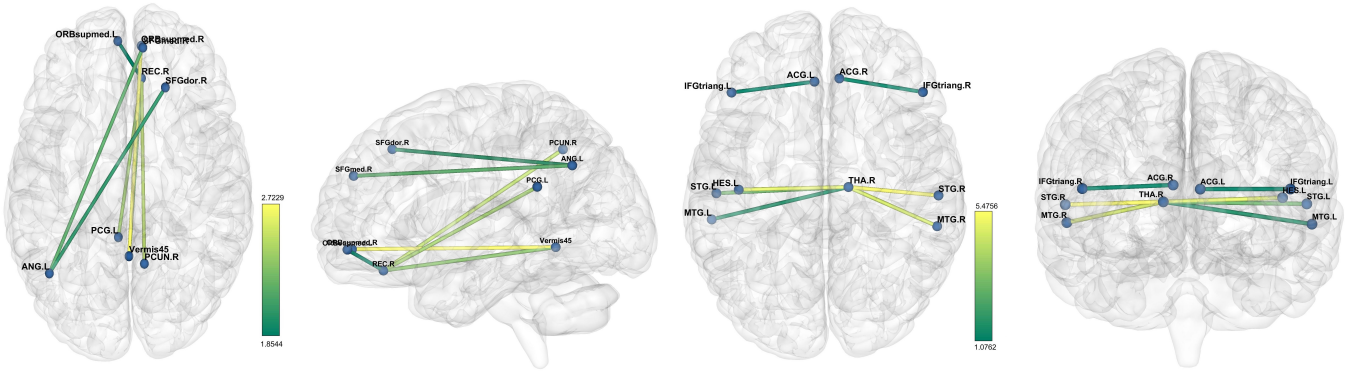


Fig. 3: Experimental results including accuracies and training times for each of the studied approaches.

a fair chance at performing well. It was also ensured that the parameters listed in their paper were included in the grid search.

All of the experimental results can be seen in the provided replication package. Figure 3 provides the average prediction accuracies (with standard deviations) as well as the average training runtimes for the nested and non-nested cross validation experiments. Several observations can immediately be made after inspecting these charts:

1) None of the accuracies are particularly high in the context of a medical diagnosis. Because the class distribution is approximately half and half, a simple classification model that always predicts a certain class would achieve about 50% accuracy. These results are clearly above that (for the most part), meaning some differences in the classes are certainly detectable when using fMRI data, but they are not accurate enough for an expert in the field to trust their predictions.
2) The approaches of Lanciano *et al.* and Iidaka do not achieve the results claimed by the respective authors.



(a) TD: Axial (b) TD: Sagittal (c) ASD: Axial (d) ASD: Coronial

Fig. 4: The most discriminative edges of the corresponding classes. The range of edge weights is indicated by the scales for each class. Edge weights represent the absolute value of the respective sum of the difference network edges when each edge was selected during the 50 test folds.

3) No single approach outperforms all others in every experiment with respect to accuracy.
4) The accuracies of the approaches using the thresholded correlation matrices are generally lower than those using raw correlation matrices as inputs.
5) DE consistently outperforms all other approaches in terms of training runtimes.
6) The approaches involving the SDP solver take a significant amount of extra computational time with little return in terms of accuracy.

As mentioned previously, the parameter grids used for each approach are documented along with the precise values of the results in the *experiments* directory of this study's repository[5].

An experiment was also conducted to visualize the connections deemed most important by the DE approach. The entire group of raw correlation matrices was divided into 5 folds to simulate 5-fold cross-validation, then the 5 most positive and negative connections in the difference network for each fold were selected according to the DE algorithm and were accumulated into an adjacency matrix for each class (i.e. ASD and TD). This process was iterated 10 times to determine which edges would be deemed most important by the DE algorithm for the positive class (ASD in this case) and the negative class (TD in this case). The accumulated adjacency matrices represented the discriminative power of each edge that was chosen at least once in the 50 folds. The adjacency matrices were then thresholded to show only the most discriminative edges. These connections were visualized with the BrainNet Viewer [40] as seen in Figure 4.

VII. Discussion and Conclusions

The problem of identifying and diagnosing Autism Spectrum Disorder through resting state fMRI data alone is difficult. The highest classification accuracies reported on the ABIDE I dataset are not adequate to be trusted in practice in most fields, let alone the medical field.

[5]https://github.com/keanelekenns/brain-network-classification

The current definition of ASD may be a source of difficulty in this problem. The disorder varies widely in severity and expression. This seems to indicate that a binary diagnosis is not sufficient, and that training data should at least include the severity levels associated with ASD diagnoses, though it would also likely be useful to include specific symptom expressions or even develop data-driven definitions of subcategories within ASD to aid in classification and treatment.

Some challenges were also discovered while attempting to replicate the work of others. It was identified that this field of research would benefit from studies that emphasize reproducibility, not just explanations of methodologies, but providing software and resources to quickly and easily recreate the experiments and results described. This will lend credibility to the research done and lead to the earlier adoption of new techniques and tools.

Most prominently, however, the problem suffers from a lack of data availability. It is difficult to determine whether the low accuracies of the approaches experimented with are caused solely by the approaches themselves or by a lack of data. Obtaining more data could potentially overcome the multitude of confounding factors that come as a result of the uniqueness of each individual's brain depending on their environment, age, sex, or other factors.

Unfortunately, obtaining such data is expensive and time-consuming with the current state of brain imaging technologies such as fMRI, and the preprocessing for such data requires expertise that is not common among researchers with backgrounds in machine learning and artificial intelligence.

Furthermore, identifying ASD in children under 2 years old, who typically cannot be diagnosed behaviourally, poses an even bigger challenge due to additional difficulties in obtaining data for younger subjects.

The easiest way to alleviate this issue in the short term is to perform similar preprocessing on the ABIDE II dataset as the Preprocessed Connectomes Project has done for ABIDE I to make it more accessible to the wider research community,

though this does not help in the area of providing more data for younger subjects, and more data is certainly needed for subjects of all ages.

The work done to provide earlier, more reliable and accurate ASD diagnoses using brain imaging data will advance our understanding of ASD and improve the quality of life of many members of society. It is hoped that our findings and suggestions will aid in future efforts to meet this goal.

REFERENCES

[1] D. Martino *et al.*, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry*, vol. 19, no. 6, pp. 659–67, 2014.

[2] P. G. Enticott, H. A. Kennedy, N. J. Rinehart, B. J. Tonge, J. L. Bradshaw, J. R. Taffe, Z. J. Daskalakis, and P. B. Fitzgerald, "Mirror neuron activity associated with social impairments but not age in autism spectrum disorder," *Biological psychiatry (1969)*, vol. 71, no. 5, pp. 427–433, 2012.

[3] L. M. Hernandez, J. D. Rudie, S. A. Green, S. Bookheimer, and M. Dapretto, "Neural signatures of autism spectrum disorders: insights into brain network dynamics," *Neuropsychopharmacology (New York, N.Y.)*, vol. 40, no. 1, pp. 171–189, 2015.

[4] L. de la Torre-Ubieta, H. Won, J. L. Stein, and D. H. Geschwind, "Advancing the understanding of autism disease mechanisms through genetics," *Nature medicine*, vol. 22, no. 4, pp. 345–361, 2016.

[5] APA, *Diagnostic and statistical manual of mental disorders : DSM-5™*. Washington, DC ;: American Psychiatric Publishing, a division of American Psychiatric Association, 5th edition. ed., 2013.

[6] R. Grant and M. Nozyce, "Proposed changes to the american psychiatric association diagnostic criteria for autism spectrum disorder: Implications for young children and their families," *Maternal and child health journal*, vol. 17, no. 4, pp. 586–592, 2013.

[7] J. Zeidan, E. Fombonne, J. Scorah, A. Ibrahim, M. S. Durkin, S. Saxena, A. Yusuf, A. Shih, and M. Elsabbagh, "Global prevalence of autism: A systematic review update," *Autism research*, vol. 15, no. 5, pp. 778–790, 2022.

[8] C. Lord, S. Risi, P. S. DiLavore, C. Shulman, A. Thurm, and A. Pickles, "Autism From 2 to 9 Years of Age," *Archives of General Psychiatry*, vol. 63, pp. 694–701, 06 2006.

[9] S. L. Hyman, S. E. Levy, S. M. Myers, D. Z. Kuo, S. Apkon, L. F. Davidson, K. A. Ellerbeck, J. E. Foster, G. H. Noritz, M. O. Leppert, *et al.*, "Identification, evaluation, and management of children with autism spectrum disorder," *Pediatrics*, vol. 145, no. 1, 2020.

[10] C. P. Santana, E. A. de Carvalho, I. D. Rodrigues, G. S. Bastos, A. D. de Souza, and L. L. de Brito, "rs-fmri and machine learning for asd diagnosis: a systematic review and meta-analysis," *Scientific reports*, vol. 12, no. 1, pp. 6030–6030, 2022.

[11] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *NeuroImage (Orlando, Fla.)*, vol. 145, no. Pt B, pp. 137–165, 2017.

[12] C.-W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager, "Building better biomarkers: brain models in translational neuroimaging," *Nature neuroscience*, vol. 20, no. 3, pp. 365–377, 2017.

[13] J. O. Maximo, E. J. Cadena, and R. K. Kana, "The implications of brain connectivity in the neuropsychology of autism," *Neuropsychology review*, vol. 24, no. 1, pp. 16–31, 2014.

[14] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "When will ai exceed human performance? evidence from ai experts," *Journal of Artificial Intelligence Research*, vol. 62, pp. 729–754, 2018.

[15] A. L. Fogel and J. C. Kvedar, "Artificial intelligence powers digital medicine," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–4, 2018.

[16] M. A. Brzezicki, N. E. Bridger, M. D. Kobetić, M. Ostrowski, W. Grabowski, S. S. Gill, and S. Neumann, "Artificial intelligence outperforms human students in conducting neurosurgical audits," *Clinical Neurology and Neurosurgery*, vol. 192, p. 105732, 2020.

[17] A. B. Kahng, "Ai system outperforms humans in designing floorplans for microchips," 2021.

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

[19] C. Elkan, "Evaluating classifiers," *San Diego: University of California*, 2012.

[20] A. Hassan, R. Sulaiman, M. Abdulgabber, and H. Kahtan, "Towards user-centric explanations for explainable models: A review," *Journal of Information System and Technology Management*, vol. 6, pp. 36–50, 09 2021.

[21] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, 2021.

[22] O. Asan, A. E. Bayrak, and A. Choudhury, "Artificial intelligence and human trust in healthcare: Focus on clinicians," *J Med Internet Res*, vol. 22, p. e15154, Jun 2020.

[23] T. Lanciano, F. Bonchi, and A. Gionis, *Explainable Classification of Brain Networks via Contrast Subgraphs*, p. 3308–3318. New York, NY, USA: Association for Computing Machinery, 2020.

[24] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowledge-Based Systems*, vol. 151, pp. 78–94, 2018.

[25] T. Iidaka, "Resting state functional magnetic resonance imaging and neural network classified autism and control," *Cortex*, vol. 63, pp. 55–67, 2015.

[26] H. S. Nogay and H. Adeli, "Machine learning (ml) for the diagnosis of autism spectrum disorder (asd) using brain imaging," *Reviews in the neurosciences*, vol. 31, no. 8, pp. 825–841, 2020.

[27] M. Khodatars, A. Shoeibi, D. Sadeghi, N. Ghaasemi, M. Jafari, P. Moridian, A. Khadem, R. Alizadehsani, A. Zare, Y. Kong, A. Khosravi, S. Nahavandi, S. Hussain, U. R. Acharya, and M. Berk, "Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: A review," *Computers in biology and medicine*, vol. 139, pp. 104949–104949, 2021.

[28] Y. Liu, L. Xu, J. Li, J. Yu, and X. Yu, "Attentional connectivity-based prediction of autism using heterogeneous rs-fmri data from cc200 atlas," *Experimental neurobiology*, vol. 29, no. 1, pp. 27–37, 2020.

[29] F. Z. Subah, K. Deb, P. K. Dhar, and T. Koshiba, "A deep learning approach to predict autism spectrum disorder using multisite resting-state fmri," *Applied Sciences*, vol. 11, no. 8, 2021.

[30] 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, 2017.

[31] H. Sewani and R. Kashef, "An autoencoder-based deep learning classifier for efficient diagnosis of autism," *Children*, vol. 7, no. 10, 2020.

[32] R. M. Thomas, S. Gallo, L. Cerliani, P. Zhutovsky, A. El-Gazzar, and G. van Wingen, "Classifying autism spectrum disorder using the temporal statistics of resting-state functional mri data with 3d convolutional neural networks," *Frontiers in Psychiatry*, vol. 11, 2020.

[33] 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, pp. 599–599, 2013.

[34] T. M. Epalle, Y. Song, Z. Liu, and H. Lu, "Multi-atlas classification of autism spectrum disorder with hinge loss trained deep architectures: Abide i results," *Applied soft computing*, vol. 107, pp. 107375–, 2021.

[35] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29–52, 2022.

[36] A. Perotti, P. Bajardi, F. Bonchi, and A. Panisson, "Graphshap: Motif-based explanations for black-box graph classifiers," *arXiv preprint arXiv:2202.08815*, 2022.

[37] C. Abrate and F. Bonchi, "Counterfactual graphs for explainable classification of brain networks," 2021.

[38] J. Cadena, A. K. Vullikanti, and C. C. Aggarwal, "On dense subgraphs in signed network streams," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 51–60, 2016.

[39] C. Tsourakakis, F. Bonchi, A. Gionis, F. Gullo, and M. Tsiarli, "Denser than the densest subgraph: Extracting optimal quasi-cliques with quality guarantees," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, (New York, NY, USA), p. 104–112, Association for Computing Machinery, 2013.

[40] M. Xia, J. Wang, and Y. He, "Brainnet viewer: a network visualization tool for human brain connectomics," *PloS one*, vol. 8, no. 7, pp. e68910–e68910, 2013.