Identifying Autism Spectrum Disorder in fMRI Brain Scans

by

Keanelek Enns
B.Sc., University of Victoria, 2021

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Identifying Autism Spectrum Disorder in fMRI Brain Scans

by

Keanelek Enns
B.Sc., University of Victoria, 2021

Supervisory Committee

Dr. Alex Thomo, Supervisor
(Department of Computer Science)

Dr. Venkatesh Srinivasan, Supervisor
(Department of Computer Science)

Supervisory Committee

Dr. Alex Thomo, Supervisor
(Department of Computer Science)

Dr. Venkatesh Srinivasan, Supervisor
(Department of Computer Science)

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 the work of Lanciano *et al.* in their paper "Explainable Classification of Brain Networks via Contrast Subgraphs", but found inconsistent results 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. A replication 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 replicability in studies.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

My beautiful wife, Emily, whom I love with all my heart, and who makes every part of life better.

My entire family, for encouraging me and loving me unconditionally (especially my mom, who has always been my biggest cheerleader).

My good friend Jacob, for regularly checking in on me and celebrating my victories with me.

My supervisors, Alex and Venkatesh, for their support, encouragement, and mentorship.

You keep him in perfect peace whose mind is stayed on you, because he trusts in you.
Trust in the Lord forever, for the Lord God is an everlasting rock.
Isaiah 26:3-4 ESV

DEDICATION

I dedicate this thesis to Jesus Christ, my God and my Refuge through all of life's storms. He is good beyond measure and truly faithful in all circumstances.

Chapter 1

Introduction

The American Psychological Association (APA) defines Autism Spectrum Disorder (ASD) in the following way:

> *"Any one of a group of disorders with an onset typically occurring during the preschool years and characterized by varying but often marked difficulties in communication and social interaction. ASD was formerly said to include such disorders as the prototype autism, Asperger's disorder, childhood disintegrative disorder, and Rett syndrome; it was synonymous with pervasive developmental disorder but more commonly used, given its reflection of symptom overlap among the disorders. It is now the official term used in DSM–5, where it encompasses and subsumes these disorders: Autism, Asperger's disorder, and childhood disintegrative disorder are no longer considered distinct diagnoses, and medical or genetic disorders that may be associated with ASD, such as Rett's syndrome, are identified only as specifiers of the disorder." [3]*

The disorder encompasses a wide variety of conditions and is expressed in many ways. The criteria for diagnosing ASD, as given by the DSM-5, are not as concrete compared to other neurological disorders [4] and even underwent significant changes in 2013 [5] as alluded to in the above definition. The causes of ASD, as well as the disorder itself, are not well understood by the scientific and medical communities [6, 7, 8]. Moreover, the disorder is common and affects a large portion of the population. In 2021, the CDC published a study that found 1 in 44 children from a 2018 sample of 8-year-olds were diagnosed with ASD [9]. It is estimated that about 1% of children are diagnosed with ASD globally, with it being nearly four times more common in

males, and it has been observed that estimates have increased over time and vary widely between sociodemographic groups [10].

ASD is currently diagnosed by observing behaviour [4] and cannot reliably be diagnosed until an individual reaches about 2 years of age [11]. In many cases, a diagnosis is not given until much later. However, 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 [12]. If ASD could be diagnosed earlier, and more objectively, it could help prepare families, ensure the individual has the necessary support and understanding, and allow early intervention to improve social skills.

Attempting to diagnose ASD by observing behaviour in children younger than 18 months (which is considered to be about the earliest it can currently be diagnosed) is unlikely to yield useful results. However, as ASD is a neurodevelopmental disorder, it would make sense to look at the brain when attempting to identify ASD before behavioural traits are apparent (although attempts have also been made to investigate other areas of the body [13]). 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 physically, 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 [14, 15, 16, 17]. 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 outperform humans significantly in a multitude of domains [18, 19, 20, 21], and the domain of graph classification is no exception [22]. But how is performance measured

for graph classification? Metrics such as accuracy, precision, and recall are essential for evaluating any classifier [23], 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 [24, 25].

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 [26]. 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 [1]. This translation of a graph into a more understandable representation is known as a graph embedding and is fundamental to many graph classification problems [27]. The graph embedding employed by Lanciano *et al.* involves thresholding correlation values of constructed brain networks and the use of contrast subgraphs (CSs), which are defined later in this thesis.

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 outline the research questions asked by this study:

RQ1 Can this study recreate the results of Lanciano *et al.*'s work?

RQ2 Can their approach be improved upon in terms of accuracy, performance, explainability, and simplicity?

RQ3 How does the thresholding of correlation matrices representing brain networks affect the accuracy of classifiers in this context?

The work done to answer these questions resulted in the following contributions to this area of research:

- A replication of the novel work done by Lanciano *et al.* was performed.
- Various modifications to the CS method were implemented.

- A new approach to the problem was introduced named Discriminative Edges (DE) which provides a simpler solution with a fraction of the computational complexity of the CS method.
- An effect size thresholding approach was implemented for comparison, though not as a full replication [28].
- A replication package was provided to recreate all the work done in this study including the replication of Lanciano *et al.*'s work and the effect size thresholding adaptation.
- Suggestions were made for the future of this area of research.

The rest of this thesis is organized as follows:

Chapter 2 gives the necessary background information to understand the work done in this study and concludes with a discussion of related works.

Chapter 3 describes the approaches used to solve the problem that were studied in this thesis.

Chapter 4 summarizes the experiments undertaken in this study and presents their results.

Chapter 5 includes a discussion and analysis of the results obtained in this study, along with observations about the future of this research area.

Chapter 6 states the main conclusions of this study, answers the proposed research questions, and provides suggestions for future work.

Chapter 2

Background

ASD is a neurodevelopmental disorder that varies widely in expression and severity, it affects many individuals worldwide, and its impact on society is significant. The problem of identifying ASD through biological information such as brain scans has been studied heavily over recent decades as it is crucial to gain a better understanding of its biological features so that earlier diagnoses can be given and better treatments can be developed to improve the quality of life for those affected by it.

This chapter will provide context for the work done in this study, equip the reader to understand the experiments conducted, and discuss related works.

2.1 fMRI Data and Preprocessing

Though many methods exist for studying the brain such as MEG, EEG, and CT scans, a prominent method for this investigation has been the analysis of functional magnetic resonance imaging (fMRI) data, which emphasizes the activity levels of ROIs of the brain and how they communicate with one another.

fMRI works by measuring blood-oxygen-level-dependent (BOLD) signals in the brain. When ROIs in the brain are active, the body increases the flow of oxygenated blood to the neurons in the region. The extent to which this occurs can be measured using magnetic resonance [29]. To summarize, higher BOLD signals in a region of the brain correspond to greater levels of activity in that region.

The temporal resolution (TR) of the scan is the time between samples of a subject's brain. This is usually a few seconds in length but can vary depending on the study [30]. A three-dimensional image of BOLD intensities is outputted for each scan

of the subject's brain. The brain is scanned many times during a session, resulting in a group of three-dimensional images that show the fluctuation of the BOLD signal in the brain over time. This four-dimensional data is referred to as a group of time series in this thesis.

The granularity of the BOLD signal measurements is determined by voxels, which are three-dimensional pixels commonly on the scale of cubed millimetres. These voxels each correspond to a time series as they report a BOLD intensity value throughout the fMRI scans and can be grouped together to compose ROIs in the brain.

The grouping and labelling of such voxels are done using a brain atlas. Brain atlases define ROIs in a three-dimensional frame of reference and can allow for comparisons between individuals as well as provide useful logical groupings of signals from the fMRI scans [31]. To account for variability in brain shape and size, nonlinear transformations are used to conform the brains of subjects to a common reference, but the details of this process are out of the scope of this study.

There are various sources of noise in fMRI scans such as thermal noise or noise introduced by the scanner's hardware. However, the largest source of noise is physiological noise. This includes changes in a subject's blood flow rate, blood flow volume, and oxygen usage due to their heartbeat and breathing (along with other factors). It also includes any movement from the subject during the scan, which must be corrected for to realign the brain in the scans over time [30].

Various further processing steps are often performed, some of which are optional or even controversial, such as band-pass filtering or global signal regression (GSR) which removes the average time series values of all brain voxels of a subject [32]. After the preprocessing is complete, what remains is a time series of BOLD signals for each defined ROI of the brain.

2.1.1 The ABIDE Dataset and the PCP

In August 2012, the Autism Brain Imaging Data Exchange (ABIDE) released data for their first initiative known as ABIDE I (referred to simply as the ABIDE dataset from here on) [2]. This initiative was coordinated across 17 international sites to collect and share resting state fMRI (R-fMRI) data with the wider research community. The dataset includes R-fMRI scans from 1112 individuals: 539 were diagnosed with ASD and 573 were typically developed (TD) controls. The dataset also includes phenotypic data about the individuals with information such as each subject's age,

sex, and whether the subject's eyes were closed during the scan.

This dataset has supported the significant surge of research on this specific topic as well as other tangentially related areas of research. It has also allowed for cross-site comparisons to see the impacts of different study environments and procedures. This study uses the ABIDE dataset exclusively in its experiments and analysis.

To make the dataset more accessible, and thereby increase its impact, the Preprocessed Connectomes Project (PCP) set out to preprocess the ABIDE dataset in a variety of ways [33]. Multiple teams used various tools and pipelines to preprocess the R-fMRI data such that researchers without the necessary expertise or time to perform such preprocessing could simply use the datasets outputted from the PCP. Versions of the ABIDE dataset with and without band-pass filtering and GSR were released from each preprocessing pipeline to allow researchers flexibility based on the requirements of their studies. This initiative also helped to produce closer comparisons between studies that used the same preprocessed data.

2.2 BOLD Timeseries to Correlation Matrices

BOLD time series, derived from the ABIDE dataset by the PCP, serve as the common input for each of the approaches implemented in this study. However, studying the activity of ROIs independently can only provide so much useful information. It is known that the brain works as a complex network of neurons where each region communicates with other regions to produce thoughts and actions.

Figure 2.1 displays the time series for three ROIs in a single subject's brain. The BOLD signal of each ROI is seen fluctuating throughout the scan. Notice the high correlation between the activity of the blue and orange time series in contrast to the lack of correlation between them and the green time series.

The exact nature of how regions of the brain communicate and how that would be reflected in their BOLD signal time series is outside the scope of this research. However, it is generally accepted that functional connections between ROIs can be represented by how correlated their BOLD time series are [2]. When two regions consistently show similar BOLD signal activity, they are said to be functionally connected.

Although there are alternatives, the extent to which two collections of data points are correlated is often calculated using the Pearson correlation coefficient. All of the approaches discussed in this thesis calculate the pairwise Pearson correlation

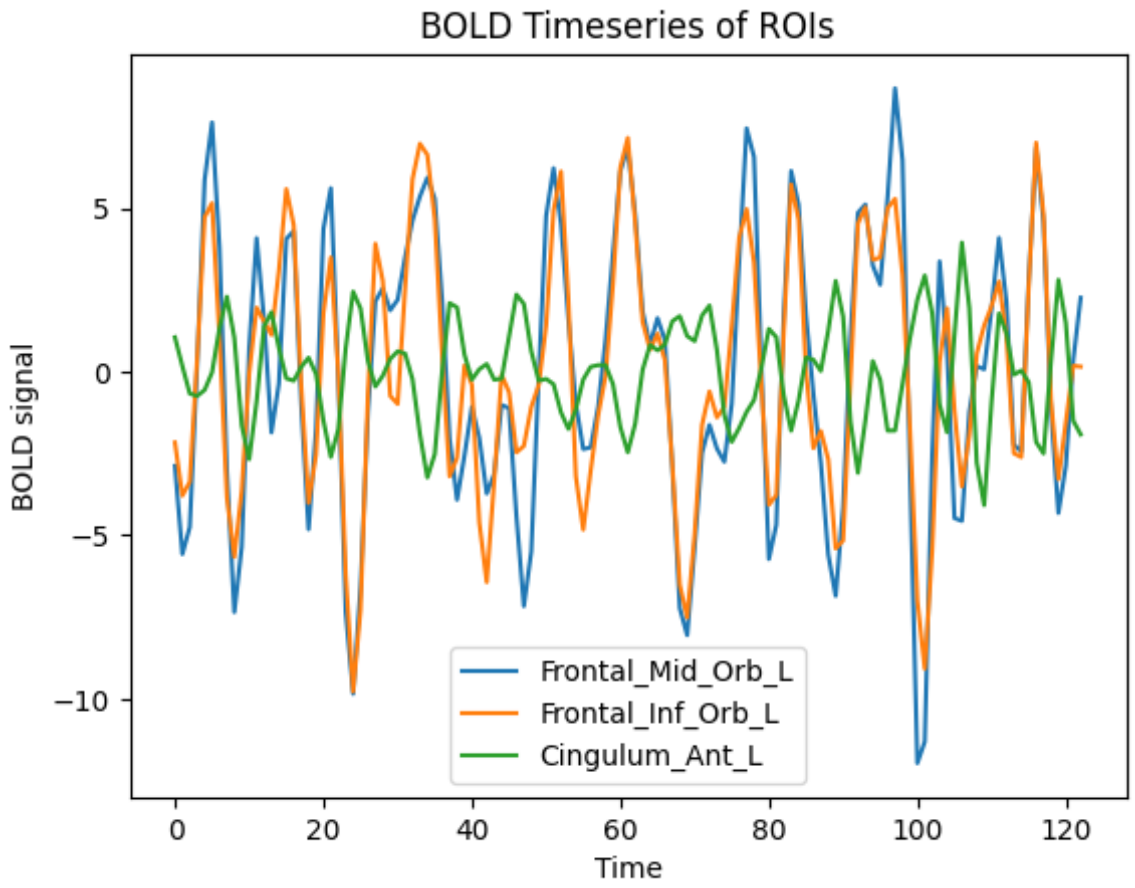


Figure 2.1: BOLD time series of three ROIs for subject 50794 from the ABIDE dataset [2]. The ROIs are labelled according to the AAL label file provided by the PCP. The orbital region of the left middle frontal gyrus and the orbital region of the left inferior frontal gyrus are highly correlated, whereas the left anterior cingulate and paracingulate region is negatively correlated to the others.

coefficients between each of the BOLD time series of a subject's ROIs.

Figure 2.2, shows a correlation matrix that resulted from these pairwise calculations for one subject. At this point, one can think of a correlation matrix as the weighted adjacency matrix for a brain network. That is to say, the ROIs represent nodes of the brain network or graph, and the correlation coefficients represent weighted edges in the network. Higher weighted edges correspond to stronger connections between regions in the brain. This is a critical way of thinking about these correlation matrices moving forward.

In some approaches, the correlation coefficients are given directly to various ML or deep learning (DL) models as inputs. In other approaches, higher-level features

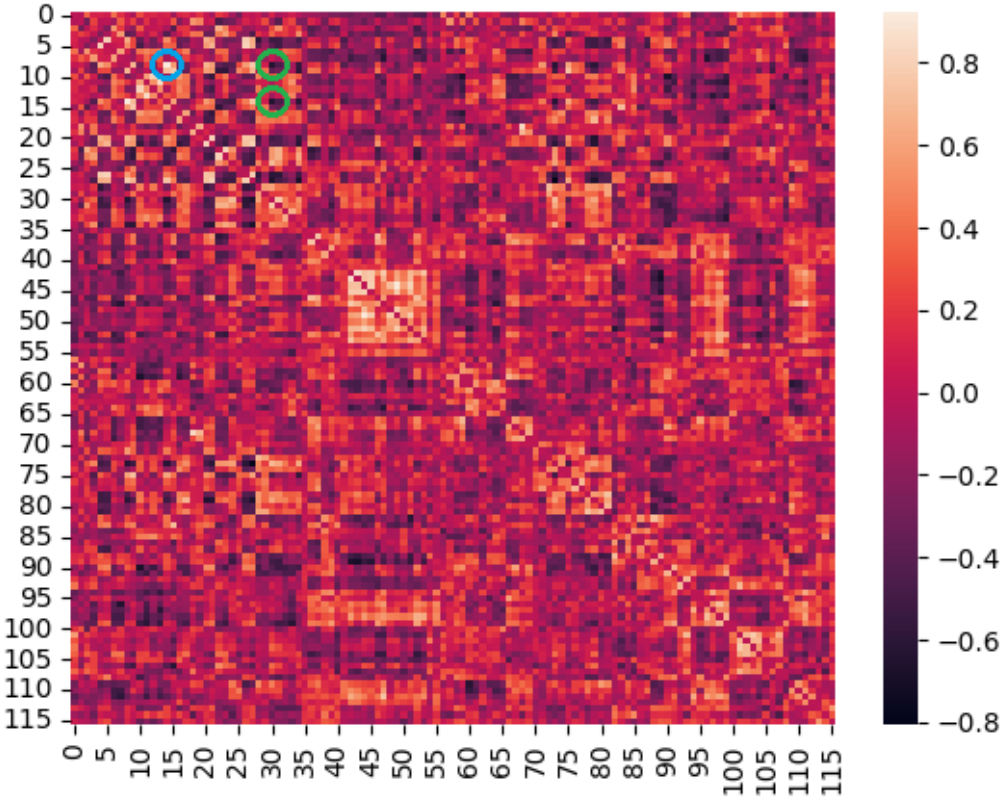


Figure 2.2: The correlation matrix of the subject in Figure 2.1. The blue circle indicates the high correlation value between the correlated ROIs, whereas the green circles indicate the negative correlation values between the third ROI and the others. The matrix represents the pairwise Pearson correlation coefficients of every ROI in the brain. The matrix is symmetric, and the main diagonal contains zeros as the correlation of each ROI to itself is irrelevant.

are derived or extracted from these correlation matrices and used for classification.

2.3 Feature Selection

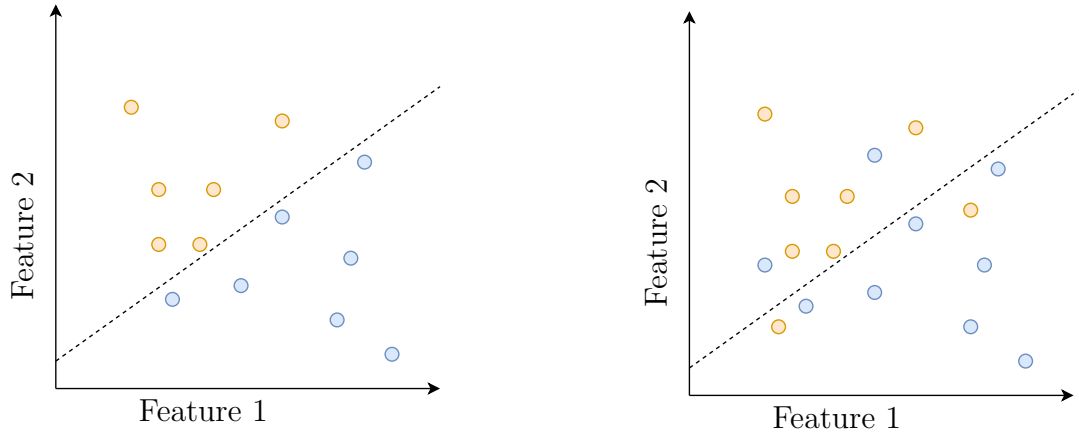
When classifying data with ML or AI, it is often necessary to translate input data into feature vectors. Feature vectors hold a fixed number of numerical values that represent relevant attributes or characteristics of the corresponding input data. A relevant attribute can be used to discriminate between the classes being predicted. This translation of input data into feature vectors positions the data in an n-dimensional space from which point many classical ML models can employ algorithms to create a "decision-boundary" which defines how the ML model will predict new input data.

The problem of creating the decision boundary given a set of data points has been studied extensively and is solved by a multitude of popular ML models. The more interesting and relevant problem in the context of classifying subjects with ASD is to determine what features and how many of them should be used to represent each subject's brain scans. There are three main aspects to consider when selecting features:

1. Separability - The features should be able to discriminate between inputs of each class. That is to say, the points of one class should be far away from the points of another class. For example, it may be reasonable to use IQ as a feature to predict whether subjects will pass a given academic test, whereas using the subjects' heights as a feature would likely not separate the subjects who pass or fail the test and would therefore provide poor predictive power.

2. Extensibility - Another term for extensibility is robustness. This relates to how applicable the features are to new data. For example, suppose Figure 2.3a represents a group of subjects used for training from which Features 1 and 2 were derived and a decision boundary was determined. Then suppose Figure 2.3b represents a new group of subjects whose classes are unknown to the model. The model misclassifies some of the subjects based on the decision boundary because the features did not appropriately account for the new data. It should be noted that a given set of features can provide great separability in training and not be extensible to new data, especially when the features of the embeddings are derived from the training data. In some cases both the training data and unseen data can be separated, but in different directions such that the model still performs poorly.

3. Explainability - This refers to how easily a human can understand and interpret a model's predictions. For example, neural networks can perform very well in many domains, but it is often difficult to understand why they are making their predictions, especially when the models take thousands of features as inputs. More features can lead to more information and can result in more accurate predictions (though sometimes it introduces noise), but having fewer, meaningful features can make a model or classification technique far more explainable which is important in this area of research, as neuroscientists wish to learn more about the causes of ASD.



(a) A group of points that can be separated linearly (i.e. with a straight line or hyperplane).

(b) A group of points that cannot be separated linearly.

Figure 2.3: Data points plotted with two features. The dotted lines represent decision boundaries.

As discussed in Section 2.2, the fMRI data for each subject is translated into a correlation matrix representing a brain network. The approaches described in Chapter 3 use these brain networks as a basis and extract meaningful features from them. Once the features are extracted, a wide variety of ML or DL algorithms can be used to train models for classification.

2.4 Related Works

The problem of classifying ASD using fMRI data has been studied intensively over the past decade [14, 15, 16, 17, 34, 35]. Unsurprisingly, there are many approaches to solving the problem.

2.4.1 Diagnosis with Behavioural Information

Misman *et al.* [36] claim 99% accuracy using a DNN on an ASD dataset comprised of behavioural and family history information. This high accuracy should not come as a surprise as ASD is currently defined and diagnosed using behavioural information. The point of this research is to find an informative and reliable way to diagnose ASD using just biological markers in fMRI scans.

Abbas *et al.* sought to improve the accessibility of screening and diagnosis techniques by using a mobile application for parents to submit questionnaires and other information about their child's behaviour. They utilized the Cognoa software[1] and various ML techniques to test the viability of such a strategy for aiding in ASD diagnosis [37, 38]. Though it aims to alleviate the long processes currently required to have experts provide an ASD diagnosis, it still relies on behavioural information, which may not appear until later in the child's development as previously discussed. The same applies to any behaviourally based classification method. The goal of the current study is to work towards identifying a biologically based marker of ASD.

2.4.2 Machine Learning

Liu *et al.* conducted a fairly similar study to the present study but used the CC200 Atlas and selected features from the derived correlation matrices using the Extra Trees algorithm from the scikit-learn library[2] [39]. They claim an accuracy of 72% on the ABIDE dataset.

Most of the work in this area of research creates correlation matrices using BOLD time series. This is because the raw four-dimensional fMRI data is too large to be useful for most ML strategies. Thomas *et al.* attempted a different method for reducing the size of the data. They collapsed the temporal dimension using various metrics such as regional homogeneity. After feeding the newly aggregated three-dimensional data into a 3D Convolutional Neural Network (3D-CNN), their accuracies on the full ABIDE datasets reached about 66%, and they found that using an SVM classifier achieved similar results [40].

Deep Learning and Neural Networks

Subah *et al.* report a high accuracy of 88% on the ABIDE dataset using a Dense Neural Network (DNN) and the BASC brain atlas, however, they do not focus on explainability and simply use the entire flattened correlation matrix as input to the DNN rather than selecting features [41]. Guo *et al.* determined features using sparse auto-encoders and compared the performance of the DNN when using the selected features compared to inputting the raw correlation values; the classifier using the selected features obtained accuracies up to 9% higher [42]. Many other studies use

[1] https://cognoa.com/providers/

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

DL models for this task but focus primarily on fine-tuning the model architectures rather than feature engineering [43] or use complex feature selection techniques that are not interpretable from a human perspective [35, 42].

On the other hand, Kong *et al.* use a measure called F-score to select the top 3000 features (which are correlation values in this case). They use a DNN on the selected features and a small subset of the ABIDE dataset and claim an accuracy of 90% [22]. 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 dataset (individuals under 20 years of age) and claim about 90% accuracy [28]. Iidaka's approach will be discussed further in Section 3.3 as the feature selection component was used for comparison in this study.

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 [44]. It is also likely that over-fitting occurs in studies with small datasets; such models are not extensible to new datasets.

It is undeniable that DL techniques have been shown to achieve superior results compared to classical ML models in various domains. However, as discussed in Section 2.3, the focus of this study is to find and derive meaningful features from brain correlation matrices that can then be used for classification. While this study employs the use of basic ML models (specifically SVM) to provide consistent comparisons, the features found could just as easily be passed as inputs to various Neural Networks to achieve potentially higher accuracies.

Multi-Atlas Classification

Recall the discussion of brain atlases in Section 2.1. Many approaches, including the ones in this study, rely on a single brain atlas to abstract the raw fMRI data. However, there is no single brain atlas that is considered superior. Instead, each atlas has advantages and disadvantages in how it represents the complex organ.

In a recent study, Epalle *et al.* successfully utilized information from multiple brain atlases simultaneously to generate predictions on the ABIDE dataset [45]. 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. Ultimately, the extra information was deemed useful in providing more accurate predictions.

2.4.3 Explainable Classification

A problem with Neural Networks and DL models (often known as black-box classifiers) is that they are not very transparent or understandable when it comes to how they make their classifications. There has been a recent trend towards more explainable AI (sometimes referred to as XAI) [46] and some tools are available for explaining the predictions of such black-box models [24, 25], 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 [47], 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 [48].

The present study was heavily influenced by the work of Lanciano *et al.* as it began with a replication of their paper [1]. They claim an accuracy of 86% on a subset of the ABIDE 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.

Coupette *et al.* devised an algorithm for identifying characteristic subgraphs that show common structures between groups of graphs as well as contrastive structures. One of the examples they used to illustrate their technique was brain networks from adolescents in the ABIDE dataset [49]. This work is very similar to the work of Lanciano *et al.* as they sought to find characteristic subgraphs that were dense in one group of brain networks and sparse in the other.

Wang *et al.* recently proposed a method for embedding whole graphs in a gen-

eralizable way that does not require any hyperparameters [50]. It was based on the DHC (Degree, H-index, and Coreness) theorem and Shannon Entropy (E), abbreviated as DHC-E. The embedding has a moderate number of dimensions depending on the graph set analyzed and is based on the h-index values of each vertex in the graph. They used principal component analysis (PCA) to present the graph embeddings in two dimensions and showed that it distinguishes between different types of graphs (e.g. brain networks and random networks). Though PCA is useful for visualizations, the modified axes do not have much meaning to humans, making this approach less interpretable. The technique was also not used to distinguish between classes of brain networks, which is far more difficult than distinguishing between types of graphs.

Evidently, much work is still needed in this area. Particularly, models need to be constructed with interpretability and explainability in mind, while also obtaining high performance metrics.

Chapter 3

Approaches

This chapter will describe the approaches that were replicated, improved upon, and created during this study.

3.1 Contrast Subgraphs

In 2020, Lanciano *et al.* approached the problem of ASD classification with a focus on explainability and interpretability [1]. They set out to find subgraphs in the brain networks that could be used to discriminate between the two classes. Using these subgraphs, they represented brain networks with only two features that a human could reasonably understand. These special subgraphs are called *Contrast Subgraphs* (CSs), which is also the term this thesis uses to reference their approach.

In an attempt to reduce the noise in the input data, Lanciano *et al.* apply a threshold to each correlation matrix. They set a threshold t to be the 80^{th} percentile of the correlation coefficients. The brain network becomes unweighted, as any edge with a weight less than t is removed (i.e. assigned a value of 0), whereas edges with a weight above t remain in the network (i.e. assigned a value of 1), but are no longer weighted. Figure 3.1 shows the same correlation matrix seen in Figure 2.2, but with this thresholding procedure applied.

The notation necessary to understand the CS approach is as follows. Let the i^{th} brain network of some class (or diagnostic category) $\mathcal{A}$ be represented as an undirected, unweighted graph $G_i^{\mathcal{A}} = (V, E_i^{\mathcal{A}})$, where V is the common vertex set representing the ROIs of the brain according to some brain atlas and $E_i^{\mathcal{A}}$ is the set of edges belonging to $G_i^{\mathcal{A}}$ (note: $E_i^{\mathcal{A}} \subset V \times V$). Let a *summary graph* corresponding to a set

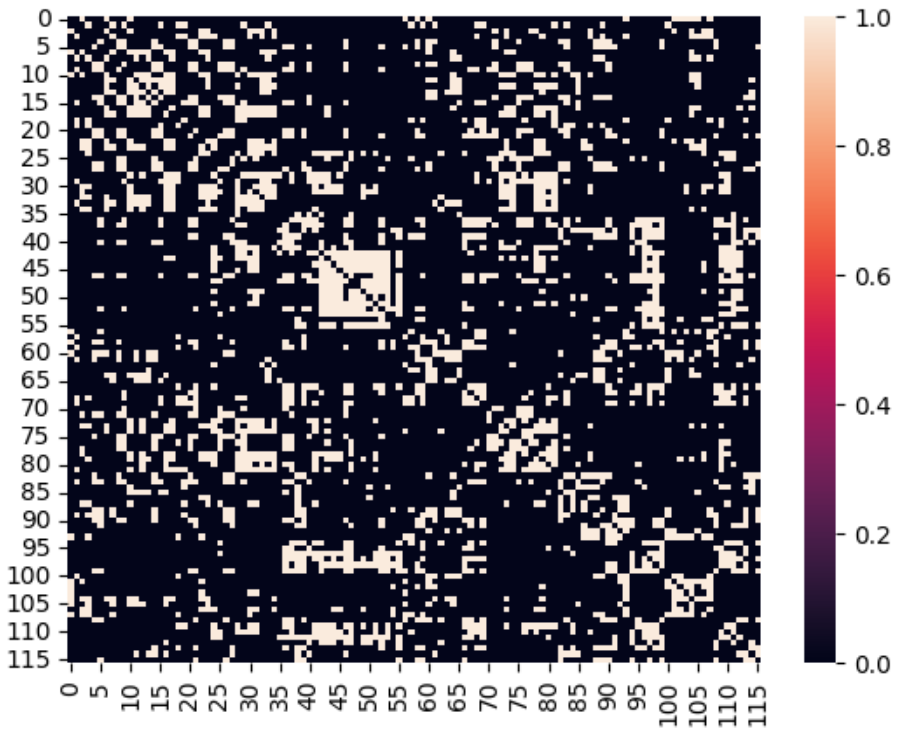


Figure 3.1: The correlation matrix of the subject in Figure 2.2, but with 80^{th} percentile thresholding applied. This can be viewed as an unweighted, undirected brain network where edges represent strong functional connections in the brain.

of brain networks in class $\mathcal{A}$ be a weighted, undirected graph $G^{\mathcal{A}} = (V, w^{\mathcal{A}})$, where $w^{\mathcal{A}} : V \times V \to \mathbb{R}_+$ is a weight function that assigns a value to each pair of vertices in V. For vertices $u, v \in V$, define $w^{\mathcal{A}}(u, v)$ to be the fraction of networks in $\mathcal{A}$ that contain the edge (u, v).

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). Lanciano *et al.* define two problem variants of the CS approach. At this point, the problems can be described.

3.1.1 Problem 1

Contrast Subgraph Problem 1 (CSP1). Given two sets of observation graphs, i.e. the condition group $\mathcal{A} = \{G_1^{\mathcal{A}}, ..., G_{|\mathcal{A}|}^{\mathcal{A}}\}$ *and the control group* $\mathcal{B} = \{G_1^{\mathcal{B}}, ..., G_{|\mathcal{B}|}^{\mathcal{B}}\}$*, and corresponding summary graphs* $G^{\mathcal{A}} = (V, w^{\mathcal{A}})$ *and* $G^{\mathcal{B}} = (V, w^{\mathcal{B}})$*, find a subset of*

vertices $S^* \subseteq V$ *that maximizes the contrast subgraph objective*

$$\delta(S) = \sum_{u,v \in S} \left(w^{\mathcal{A}}(u,v) - w^{\mathcal{B}}(u,v) - \alpha \right)$$

where $\alpha \in \mathbb{R}_+$ *is a user-defined parameter.*

The parameter α is used to penalize large CSs. It can be adjusted to vary the proportion of edges that are considered detrimental to the contrast subgraph objective.

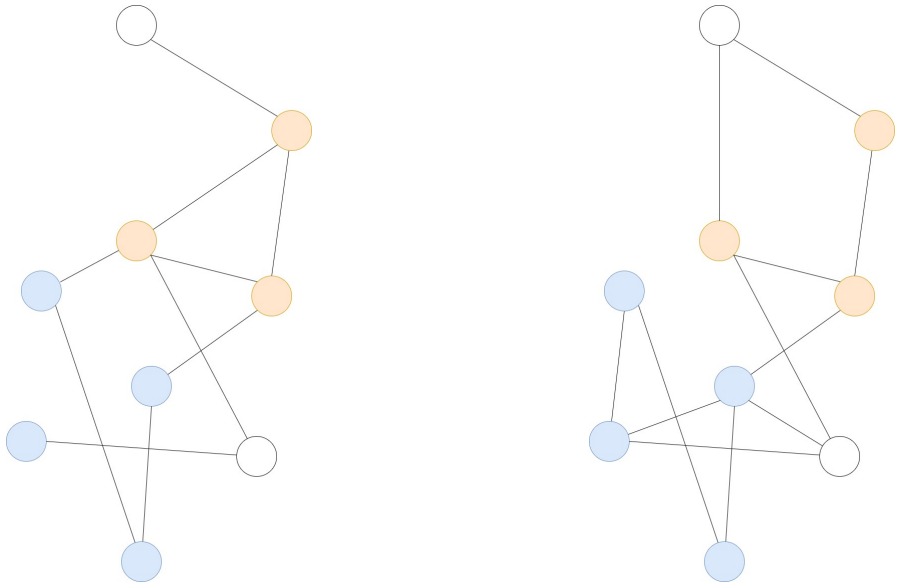
The problem can be simplified to finding a set of vertices that induce a dense subgraph in the difference between the two summary graphs. Consider the difference network $G^{\mathcal{A}-\mathcal{B}} = (V, w^{\mathcal{A}-\mathcal{B}})$, where $w^{\mathcal{A}-\mathcal{B}}(u,v) = w^{\mathcal{A}}(u,v) - w^{\mathcal{B}}(u,v), \forall u, v \in V$. In this context, finding a contrast subgraph is equivalent to finding a maximally dense subgraph in $G^{\mathcal{A}-\mathcal{B}} - \alpha$ (subtract α from each value in the difference network). There are multiple notions of density studied in the literature, but Lanciano *et al.* base the contrast subgraph objective off of the optimal quasi-clique problem [51].

The CS problem itself is based on the Generalized Optimal Quasi-Clique (GOQC) problem, which is an extension made by Cadena *et al.* for weighted, signed graphs and was proven to be NP-complete [52, 1]. Cadena *et al.* created an $O(\log n)$ approximation to the optimal solution using semidefinite programming (SDP) and a local-search procedure developed by Tsourakakis *et al.* [52, 51]. Lanciano *et al.* then repurposed this algorithm after translating the brain networks into an appropriate input as previously shown.

Recall that the original goal was to find a useful set of features to represent the brain networks of subjects such that typically developed (TD) subjects could be differentiated from those with ASD and such that the features were understandable to a human.

CSP1 is asymmetric, which is to say that the solution changes when using the difference network $G^{\mathcal{A}-\mathcal{B}}$ compared to $G^{\mathcal{B}-\mathcal{A}}$. When translating brain networks into features, Lanciano *et al.* find both CSs, one using the summary graph G^{ASD-TD} and one using G^{TD-ASD}. They then use CS overlap to create two features for each brain network.

Each feature corresponds to the number of edges in common between the brain network being translated and each CS. This is illustrated in Figure 3.2 with a toy example where the graphs represent brain networks of two different subjects. The orange nodes represent the CS found using G^{ASD-TD} and the blue nodes represent



(a) (2,3) (b) (4,2)

Figure 3.2: Two graphs representing brain networks of two different subjects. The orange nodes represent a CS found to be denser in the ASD summary graph, and the blue nodes represent a CS found to be more common in the TD summary graph. Each graph is labelled with the feature vector it would be translated into using the problem 1 variant of the CS approach.

the CS found using G^{TD-ASD}.

Figure 3.3 positions some of the actual brain networks in two dimensions based on their features for CSP1.

3.1.2 Problem 2

Lanciano *et al.* also define a symmetric variant of the CS problem, abbreviated as CSP2, which shares the same definition of CSP1, but with an alternative objective:

$$\sigma(S) = \sum_{u,v \in S} \left(|w^{\mathcal{A}}(u,v) - w^{\mathcal{B}}(u,v)| - \alpha \right)$$

In CSP2, a single CS is found in the difference network containing absolute valued edge weights (i.e. $|G^{TD-ASD}|$). The CS is used to induce subgraphs in both of the summary graphs (i.e. G^{TD} and G^{ASD}) as well as each individual brain network. The distances, computed as the L_1 norms, from the induced brain network to each induced

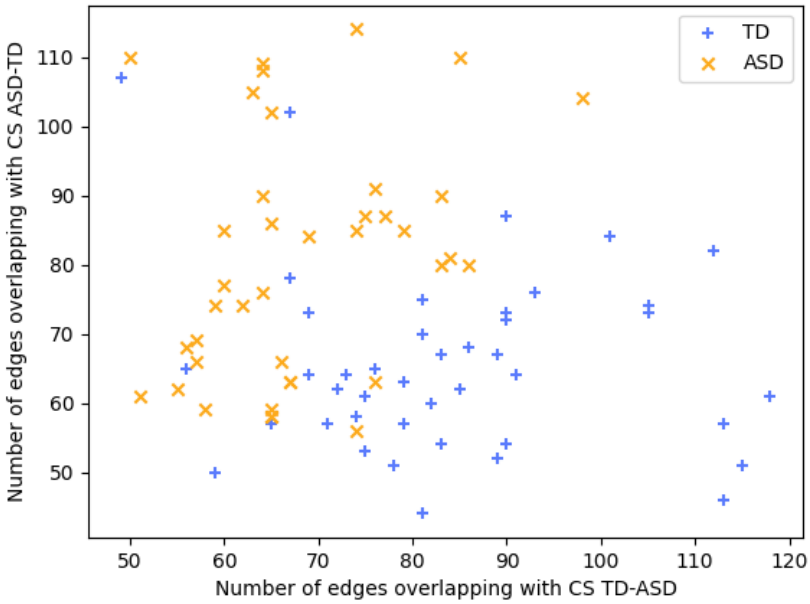


Figure 3.3: A group of brain networks plotted in two dimensions based on their features derived from the CSP1 approach. The bottom right points have more edges in common with the contrast subgraph that was found in G^{TD-ASD}, and similarly, the top left points have more edges in common with the contrast subgraph that was found in G^{ASD-TD}.

summary graph are then used as the two features for this approach.

Figure 3.4 positions some of the actual brain networks in two dimensions based on their embeddings for CSP2.

One can think of these features in the following way. If a dense subgraph is found in the absolute difference network, this means that the subgraph contains discriminative connections including both those that are more common in ASD and those that are more common in the TD class. The dense subgraph is then imposed on each of the summary graphs, so the values of these discriminative connections can be identified for each class. The features of an individual brain network are then calculated as how far the brain network is from the ASD summary graph with respect to the discriminative connections, and also how far away it is from the TD summary graph with respect to the same connections.

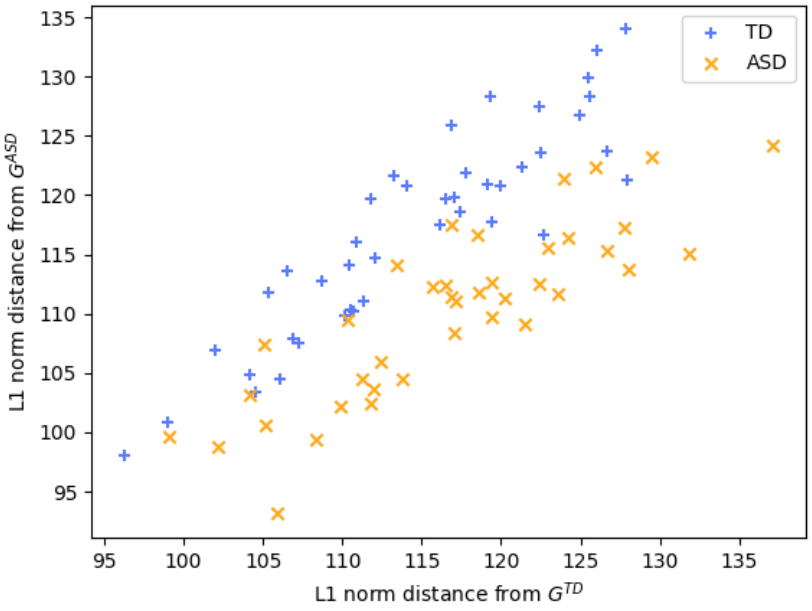


Figure 3.4: A group of brain networks plotted in two dimensions based on their features derived from the CSP2 approach. The points further to the right represent graphs with less in common with the TD summary graph (larger distance from it), and similarly, the points further to the top are graphs with less in common with the ASD summary graph.

3.1.3 Improvements

Part of the work done in this research was to replicate the study of Lanciano *et al.* Besides a general speed improvement, a few modifications were made to the original methods which are outlined in this section.

Quadratic Programming

As mentioned previously, Lanciano *et al.* utilized the work of Cadena *et al.* to find a dense subgraph in a weighted, signed network [52, 1]. Cadena *et al.* proposed the following quadratic programming (QP) formulation of the GOQC problem:

$$\text{(QP) max} \sum_{(u,v)\in E} w(u,v)\left(\frac{1+x_u x_0+x_v x_0+x_u x_v}{4}\right) - \sum_{u,v\in V, u\neq v} \alpha(u,v)\left(\frac{1+x_u x_0+x_v x_0+x_u x_v}{4}\right)$$

Subject to

$$x_0, x_u \in \{-1,1\} \text{ for all } u \in V$$

They proved that maximizing this expression is equivalent to solving the GOQC problem. They then developed an $O(\log n)$ approximation to the quadratic program with a semidefinite program and a rounding technique [52]. Lanciano *et al.* used this implementation to find a contrast subgraph in their context.

However, many tools currently exist for solving quadratic programs directly, so it seemed reasonable to utilize one such tool in this study. None of the QP solver libraries found were able to restrict potential solution vectors to the discrete set of {-1,1} (i.e. out or in) as given in the proposed quadratic program. The only satisfactory solution was to restrict each component of the solution vector to [-1,1] rounding positive values to 1 and non-positive values to -1. Though this change of constraints removes the theoretical guarantee given by Cadena *et al.* regarding the approximation to the optimal solution, the practical speed and performance increase that resulted was considerable as will be shown in Chapter 4. It is also important to remember that the main goal is not to find the densest subgraph, but to accurately classify brain networks.

After much of the experimentation had been completed, it was noticed that Cadena *et al.*'s proposed quadratic program, though equivalent to the GOQC problem, was possibly invalid with respect to the imposed constraints, which are quadratic on each component of the solution vector (i.e. $x_i^2 = 1 \,\forall\, x_i \in \mathbf{x}$ where $\mathbf{x}$ is the solution vector). It is unclear whether this constraint is valid for a quadratic program and whether it would affect the validity of the translation into an SDP problem.

In order to utilize a QP solver to identify a dense subgraph, the quadratic program had to be translated into an acceptable form. Many solvers can only minimize the expression:

$$\frac{1}{2}\mathbf{x}^T P\mathbf{x} + q^T\mathbf{x}$$

subject to:

$$G\mathbf{x} \preceq h \text{ and } A\mathbf{x} = b$$

Where $\preceq$ refers to an element-wise vector inequality.

To translate the quadratic program into this form, assume $x_0 = 1$ as in Cadena *et al.*'s proof of equivalency with the GOQC problem [52]. Additionally, assume $w(u, v)$ is defined for all $u, v \in V$ and that $w(u, v) = 0 = \alpha(u, v)$ for $u = v$ (that is to say, ignore the correlation of a ROI to itself), and let $V = \{1, 2, ..., n\}$. The expression becomes:

$$\sum_{u=1}^{n}\sum_{v=1}^{n}\left(w(u,v) - \alpha(u,v)\right)\left(\frac{1 + x_u + x_v + x_u x_v}{4}\right)$$

In the context of this study, $\alpha(u, v) = \alpha$ for all $u, v \in V, u \neq v$ and w represents the difference network of the summary graphs. Define the objective function (which comes in the form of a matrix) as $D = w - \alpha$ such that $D(u, v) = w(u, v) - \alpha(u, v)$ for all $u, v \in V$. The expression is then

$$\frac{1}{4}\left(\sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v) + \sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v)x_u + \sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v)x_v + \sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v)x_u x_v\right)$$

Note that $\sum_{u=1}^{n}\sum_{v=1}^{n} D(u, v)$ is a constant term (i.e. it does not depend on the solution vector $\mathbf{x}$), so it can be removed from the optimization problem without affecting the solution. It is also necessary to turn this into a minimization problem by multiplying by -1.

$$-\frac{1}{4}\left(\sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v)x_u + \sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v)x_v + \sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v)x_u x_v\right)$$

Moreover, notice that

$$\begin{aligned}\sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v)x_u = x_1(D(1,1) + D(1,2) + ... + D(1,n)) \\ + x_2(D(2,1) + D(2,2) + ... + D(2,n)) \\ \vdots \\ + x_n(D(n,1) + D(n,2) + ... + D(n,n))\end{aligned}$$

This is equivalent to multiplying the solution vector $\mathbf{x}$ with a vector containing the sums of each row of D. Call this vector row_sum, then the summation described becomes $row\_sum^T\mathbf{x}$ assuming all vectors are column vectors. Similarly, $\sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v)x_v = col\_sum^T\mathbf{x}$ where col_sum is a vector containing the sum of each column of D. Furthermore, because D is a symmetric matrix, $col\_sum = row\_sum$. Hence, the expression is

$$-\frac{1}{4}\left(2row\_sum^T\mathbf{x} + \sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v)x_u x_v\right)$$

Lastly, note that

$$\begin{aligned}\sum_{u=1}^{n}\sum_{v=1}^{n} D(u,v)x_u x_v = x_1(x_1D(1,1) + x_2D(1,2) + ... + x_nD(1,n)) \\ + x_2(x_1D(2,1) + x_2D(2,2) + ... + x_nD(2,n)) \\ \vdots \\ + x_n(x_1D(n,1) + x_2D(n,2) + ... + x_nD(n,n)) \\ = \mathbf{x}^T D\mathbf{x}\end{aligned}$$

and the expression simplifies to

$$-\frac{1}{2}row\_sum^T\mathbf{x} - \frac{1}{4}\mathbf{x}^T D\mathbf{x}.$$

Let $q = \frac{-row\_sum}{2}$ and $P = \frac{-D}{2}$. This gives the desired form of the expression:

$$\frac{1}{2}\mathbf{x}^T P\mathbf{x} + q^T\mathbf{x}.$$

In order to restrict the components of $\mathbf{x}$ to $[-1, 1]$ while using the required form of $G\mathbf{x} \preceq h$, there must be two constraints on each component. For each $u \in V$, impose the following two restrictions:

$$1.\, x_u \leq 1$$
$$2.\, -x_u \leq 1$$

This is done by letting h be a $2n$ by 1 vector filled with ones and by letting G be a $2n$ by n matrix defined as the identity matrixed stacked vertically with a negative identity matrix:

$$G = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \end{bmatrix}$$

After translating the problem into the correct form, the solver outputs a solution vector $\mathbf{x}$ containing values between -1 and 1. As a result of not being able to constrain the solution components to a discrete set of values, the components need to be rounded. The rounding occurs such that positive values become 1 and are included in the dense subgraph and non-positive values become -1 and are excluded from the dense subgraph.

Local Search Optimization

Due to the need for rounding, the QP and SDP solvers only approximate the densest subgraph. To improve the density of the subgraph found, a local search step is used.

In the context of this study, the density of a subgraph S is defined by $\delta(S)$ as in CSP1, and $\sigma(S)$ as in CSP2. This definition is derived from the unweighted edge-surplus $f_\alpha(S)$ defined by Tsourakakis *et al.* [51]. For the sake of simplicity, regardless of the problem being solved by the local search algorithm, this study refers to the density objective function as f (e.g. $f(S) = \delta(S)$ in CSP1).

The local search algorithm developed by Tsourakakis *et al.* and adapted by Cadena *et al.* works by considering the density of a subgraph if each vertex outside the subgraph were to be added to it and adding them when it increases the density. This is repeated until all the vertices have been considered. Afterwards, the algorithm seeks to remove a single vertex from the subgraph and does so if it finds one that, when removed, increases the density. This process is repeated until a maximum number of iterations has been reached, or no vertices can be added or taken away from the subgraph to increase its density. The algorithm finally tests to see if the complement of the optimized set of nodes obtains a higher density; if it does, it changes its output to the complement. The pseudocode for the original algorithm can be found in their paper as Algorithm 2 [51].

Considering the goal of the local search algorithm is to increase the density of the subgraph and dense subgraphs typically possess fewer nodes with many highly weighted connections, a modification to the algorithm was made such that it attempts to remove as many nodes as it can in each iteration (removing them if they retain the same density or they increase the density). This narrows the solution down to a smaller set of nodes faster. The modified local search algorithm can be seen in Algorithm 1.

Top-k Contrast Subgraphs

The final major modification follows Lanciano *et al.*'s suggestion for future work and uses the top-k contrast subgraphs. It was not implemented according to the suggested method [53], but instead, the approach described by Tsourakakis *et al.* in Section 4.1 of their paper was used [51].

The new parameter k is introduced to indicate the number of CSs that should be discovered. As each CS is discovered, the nodes of the CS are removed from the main

Algorithm 1 LocalSearch (Modified)

Input: Weighted graph $G = (V, w)$, $\alpha \in \mathbb{R}$, initial node set $S \subseteq V$, maximum number of iterations T_{MAX}

Output: Improved node set S^*

if $S = \emptyset$ **then**
    $S \leftarrow \{u, v\}$, where (u, v) is the largest edge in G
end if
$changes\_made \leftarrow TRUE$, $i \leftarrow 0$
while $changes\_made$ and $i < T_{MAX}$ **do**
    $changes\_made \leftarrow FALSE$
    $i \leftarrow i + 1$
    $found\_node \leftarrow TRUE$
    while $found\_node$ **do**
        $found\_node \leftarrow FALSE$
        if $\exists u \in V \setminus S$ s.t. $f(S \cup \{u\}) > f(S)$ **then**
            $S \leftarrow S \cup \{u\}$
            $changes\_made \leftarrow TRUE$
            $found\_node \leftarrow TRUE$
        end if
    end while
    $found\_node \leftarrow TRUE$
    while $found\_node$ **do**
        $found\_node \leftarrow FALSE$
        if $\exists u \in S$ s.t. $f(S \setminus \{u\}) \geq f(S)$ **then**
            $S \leftarrow S \setminus \{u\}$
            $changes\_made \leftarrow TRUE$
            $found\_node \leftarrow TRUE$
        end if
    end while
end while

difference network, and the next CS is discovered in the remaining network. This is repeated until all k CSs are discovered. Finally, two features are calculated for each CS found, just as in the typical approach, and the features are summed together to obtain the brain network's final two features.

3.2 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 it is to find them. 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 [1]. 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 the apparent issue with CSs. 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. What are the most important edges though? The most important edges are those that contain the most information about whether a brain network belongs to an individual with ASD or not.

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.

The top n positive edges in the difference network G^{A-B} comprise the connections in the brain that are stronger, on average, in the brains of individuals in class A, whereas the most negative n edges comprise the connections that are stronger in

class B. Because of this, one class will be referred to as the positive class, and the other, as the negative class. Note, however, that if G^{B-A} were used, the approach would function the same way, but with flipped signs.

These $2n$ discriminative edges are then used to calculate similarity scores of brain networks to the two classes based on each edge's importance, which is measured by the magnitude of the values in the difference network. This approach has been adapted for both unweighted and weighted brain networks. The details and calculations for each case are described in the sections to follow.

3.2.1 Unweighted Brain Networks

In the case of binary brain networks, like the ones derived by Lanciano *et al.*, the calculation is very simple. In the initial version of the approach, a basic dot product (i.e. element-wise multiplication followed by summation) was taken between the discriminative edges in the difference network and the binary network of an individual as illustrated by Figure 3.5.

	0	**1**	**2**	**3**	**4**	**5**	**6**
0	0	0.5	-0.1	-0.17	-0.5	0.1	-0.05
1	0.5	0	0.23	-0.45	0.2	-0.21	0.3
2	-0.1	0.23	0	-0.15	0.12	0.07	-0.34
3	-0.17	-0.45	-0.15	0	-0.23	0.4	-0.25
4	-0.5	0.2	0.12	-0.23	0	0.1	-0.11
5	0.1	-0.21	0.07	0.4	0.1	0	-0.3
6	-0.05	0.3	-0.34	-0.25	-0.11	-0.3	0

	0	**1**	**2**	**3**	**4**	**5**	**6**
0		1	1	0	0	1	0
1			0	1	1	0	1
2				1	0	1	1
3					0	1	0
4						0	1
5							1
6							

Figure 3.5: Left: An example difference network with 7 nodes. Right: An example unweighted brain network of a single individual. In this example $n = 2$. The discriminative edges for the positive and negative classes are highlighted with orange and blue respectively. The two features for the brain network would be calculated as $1 \times 0.5 + 1 \times 0.4 = 0.9$ and $1 \times -0.45 + 0 \times -0.5 = -0.45$. Only the upper triangles of the matrices are used because the given brain networks are undirected.

When a brain network possesses an edge (i.e. has a strong connection), it gains the value of the corresponding edge in the difference network. If the edge is more common in the positive or negative class (i.e. the difference network has a positive or negative value on the edge), the dot product result will be moved in the positive or negative direction respectively because the edge is multiplied by 1. However, if

the brain network does not possess an edge, it does not add the difference network's value for that edge.

It was noticed that multiplying discriminative edge weights by zero represented a loss of information. Rather than ignoring these edge weights, the values of individual brain networks could be scaled such that nonexistent edges have a value of -1 and present edges continue to have a value of 1. This means that not having an edge among the most discriminative edges counts against the similarity score for each case. For example, the modified features of the brain network in Figure 3.5 would become (0.9, 0.05) instead of (0.9, -0.45). This makes sense as if some edge is known to be common in a given class, and the currently examined brain network does not possess that edge, it would make sense to nudge the similarity score for the given brain network towards the opposite class rather than ignoring the information.

Finally, for ease of understanding, the features are made into a percentage. This is done by dividing the feature values by the sum of the corresponding discriminative edges. To continue with the example from Figure 3.5, the features would go from (0.9, 0.05) to $(\frac{0.9}{0.9} \times 100\%, \frac{0.05}{-0.95} \times 100\%) = (100\%, -5.26\%)$.

3.2.2 Weighted Brain Networks

As will be discussed in Chapter 5, the thresholding of the correlation matrices done by Lanciano *et al.* removes useful information. More successful approaches rely on the raw correlation values. For this reason, the DE approach was adapted to handle weighted brain networks (i.e. correlation matrices).

The adaptation was not as straightforward as performing the dot product between the difference network and the weighted brain network. This is for numerous reasons, an important one being the way the summary graphs, and thus the difference network, are derived in the weighted case. Recall that an edge weight in the summary graph for some class A represents the number of brain networks in that class possessing the given edge. However, in the weighted case, an edge weight in the summary graph represents the average correlation value for that edge over all the brain networks in that class.

In this case, similarity to a class is not as simple as possessing edges of one class versus another, but it is instead measured by whether the edge weights of a brain network are closer to one class or another. The difference network gives no information about the actual values of either class, just the difference between them. For example,

perhaps one class has zero correlation for an edge, whereas the other class has a very negative correlation, that edge may constitute one of the larger differences, making it an important edge, but the difference does not indicate whether the edge is positive or negative in class A or B.

Consider Figure 3.6, which illustrates the difference between the unweighted and weighted case with respect to the input graphs, the summary graphs, and the corresponding difference network. Notice how similar the difference networks are in both cases, yet the summary graphs are quite different.

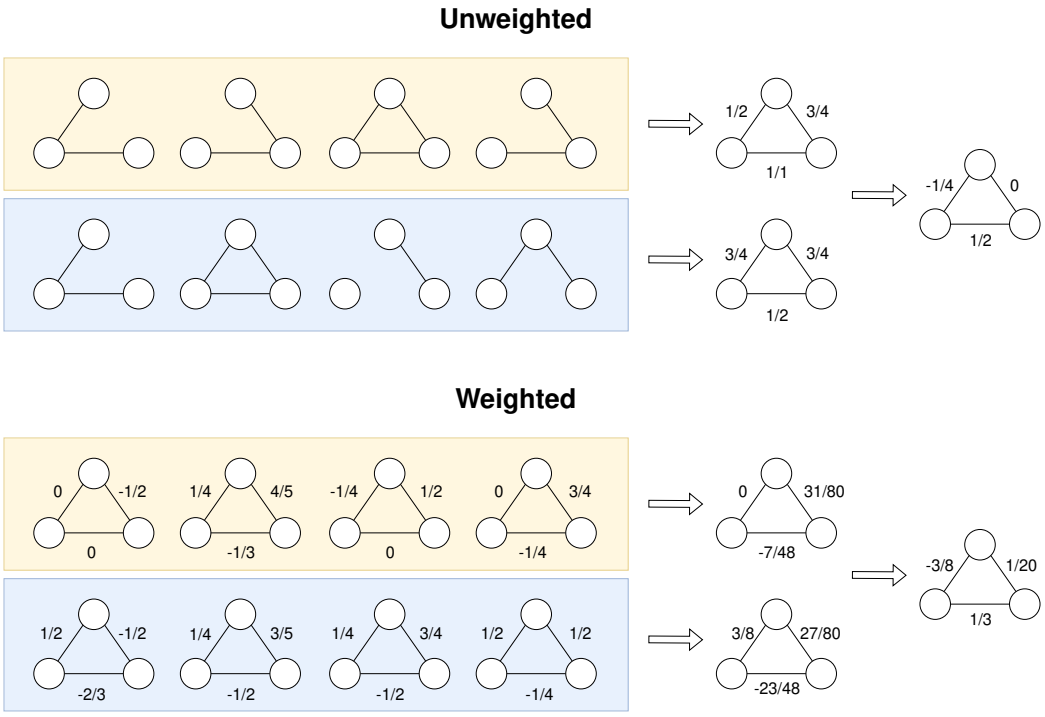


Figure 3.6: The derivation of summary graphs (middle) and difference networks (right) in the case of unweighted (top) and weighted (bottom) input graphs (left). The two classes of input graphs are denoted using the two colours.

It is useful to imagine the top or bottom n edge weights as comprising vectors in n-dimensional space. This can be seen in Figure 3.7 where $n = 2$ for simplicity with respect to visualization. The goal is still to measure the similarity of some input brain network G_i to the summary graphs of each class.

The initial idea was to use cosine similarity, which is a common measure of similarity between vectors. However, this approach did not end up performing at a satisfactory level. Instead, euclidean distances are used to measure how close a brain network's vector is to each of the vectors representing the summary graphs among

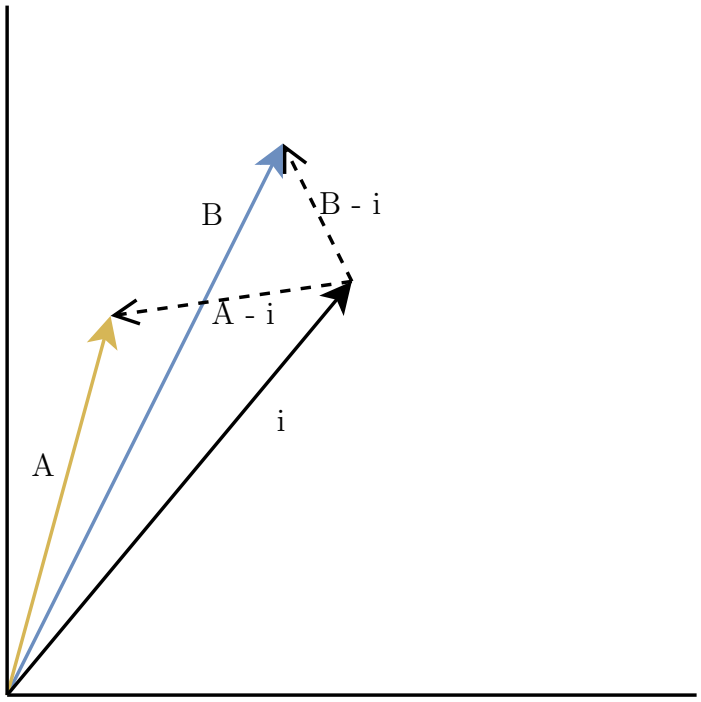


Figure 3.7: Vectors A and B represent the summary graphs of class A and B respectively. Vector i represents some input brain network. In each case, the vector can represent the whole graph (summary graph or brain network), or just the values of the top or bottom n edges. In this figure, $n = 2$, and either the top or bottom 2 edge weights are being used to vectorize each graph. The magnitude of the difference vectors can be used to determine which class vector i is more similar to, and to what extent.

the same edges. This can be seen in the difference vectors $A - i$ and $B - i$ in Figure 3.7. Note that the euclidean distance of one vector to another is equivalent to the magnitude of their difference vector.

Let A be the positive class and B be the negative class. Also, let the subscripts n^+ and n^- denote the usage of the n most positive and negative edges in the difference network respectively. Then the first feature in the weighted formulation of DE is calculated as:

$$\frac{||B_{n^+} - i_{n^+}|| - ||A_{n^+} - i_{n^+}||}{||B_{n^+} - i_{n^+}|| + ||A_{n^+} - i_{n^+}||} \times 100\% \tag{3.1}$$

This is interpreted as the percent similarity of vector i to the positive class. Some characteristics of this similarity score are listed as follows:

- When vector i is exactly the same distance from each summary vector, the

similarity score is 0%.

- When vector i is identical to class A, the similarity score is 100%.
- Conversely, when vector i is identical to class B, the similarity score is -100%.

The second feature is calculated in a very similar way, only it is calculated as the similarity of a brain network to the negative class among the most negative edges in the difference network:

$$\frac{||A_{n^-} - i_{n^-}|| - ||B_{n^-} - i_{n^-}||}{||A_{n^-} - i_{n^-}|| + ||B_{n^-} - i_{n^-}||} \times 100\% \tag{3.2}$$

This approach performed moderately well, but it was noticed that the difference network was no longer being used as a measure of importance for each edge as it was in the unweighted case. Therefore the summary graphs and each input brain network were multiplied element-wise with the difference network. This scales each graph by the same values, which leaves the calculations intact. More important edges affect the distances of the vectors more heavily rather than considering all edges to be equally important.

3.2.3 Whole Network Similarity

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. Even though these edges were not the most discriminative, it was hypothesized that using the information they provided would be more helpful than ignoring it. Therefore a third feature was derived for each case: the whole network similarity.

This feature was calculated in an almost identical way as the first feature for both the weighted and unweighted cases (see Equation 3.1 for weighted case). Only, instead of using the top n discriminative edges, every edge was used. The outcome is a measure of similarity between the input brain network and the positive class as a whole. This has the additional benefit of allowing the $2n$ discriminative edges to be compared by their relative importance.

Figure 3.8 gives a similar example as Figures 3.3 and 3.4. It shows some of the brain networks positioned in three-dimensional space using the three DE features. Note that it is more difficult to visualize three-dimensional data in the context of

this thesis, but there are many applications that could be used to visualize the three-dimensional data dynamically.

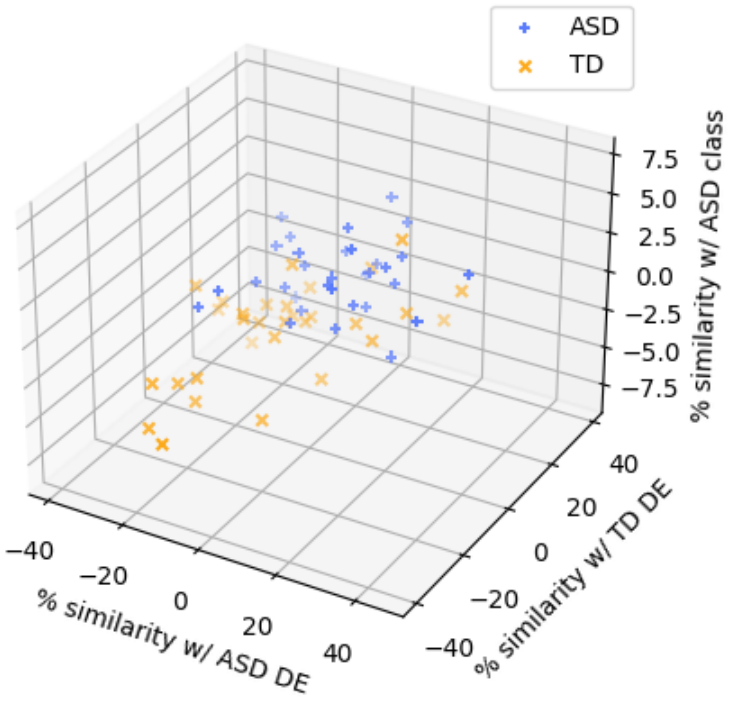


Figure 3.8: A group of brain networks plotted in three dimensions based on their features derived from the DE approach. The bottom two dimensions represent the values given by Equations 3.1 and 3.2 where ASD is the positive class and TD is the negative class.

The DE features encode more information than the contrast subgraph features, they are more efficient to compute, and they are still quite easy to visualize. This approach is also more interpretable and explainable, as the calculations are basic, and one can easily trace the importance or contribution of each edge in making a classification prediction. This does not only apply to the most influential edges in favour of the chosen class, but also those in favour of the other class.

3.3 Effect Size Thresholding

In their study from 2015, Iidaka sought to identify ASD from brain scan data using a subset of the ABIDE dataset, namely individuals under the age of 20 [28]. Iidaka computed the pairwise Pearson correlation coefficients between each of the ROI time series as described in Section 2.2.

Iidaka did not threshold the correlation matrices as in the CS approach. Instead, they utilized the raw correlation matrices as in the weighted case of the DE approach.

Iidaka normalized the correlation values for each input brain network using Fisher's Z transform (i.e. the inverse hyperbolic tangent function). They then created summary graphs in the same manner as the DE approach (i.e. take the mean of all edge weights among each class of brain networks) as well as calculated a standard deviation (SD) matrix summarizing each class.

With all of this information, Iidaka derived a matrix containing the effect size (ES) of the difference between the edges of each class. The effect size is known as Cohen's d and is given by the following formula:

$$\frac{M_A - M_B}{\sqrt{\frac{SD_A^2 + SD_B^2}{2}}} \tag{3.3}$$

where M_A and M_B are the summary graphs containing the mean edge weights of class A and B respectively and SD_A and SD_B are the matrices containing the corresponding standard deviations of the edge weights by class. Note that this formula assumes that the number of samples in class A and B are equal, which is nearly true for the ABIDE dataset, but if the inequality in sample size becomes too large, an adjustment must be made to the formula to calculate the pooled SD (i.e. denominator in this case) correctly.

The numerator of Equation 3.3 is nearly equivalent to the difference network used by DE in the weighted case. Therefore this ES matrix differs from the difference network in that its values indicate how significant the differences are between the two groups rather than simply indicating the differences. After taking the absolute value of the ESs, the resultant matrix contained importance values for each edge.

The matrix was then thresholded with the ES hyperparameter, and only edges with an effect size larger than the given threshold were used. Figure 3.9 gives an example of the chosen brain network edges in one of the cross-validation folds during experimentation. Note that the outputted feature vector has a very high number

of dimensions (often over 1000) compared to the other approaches (2 or 3) and the number of features chosen is dynamic rather than constant.

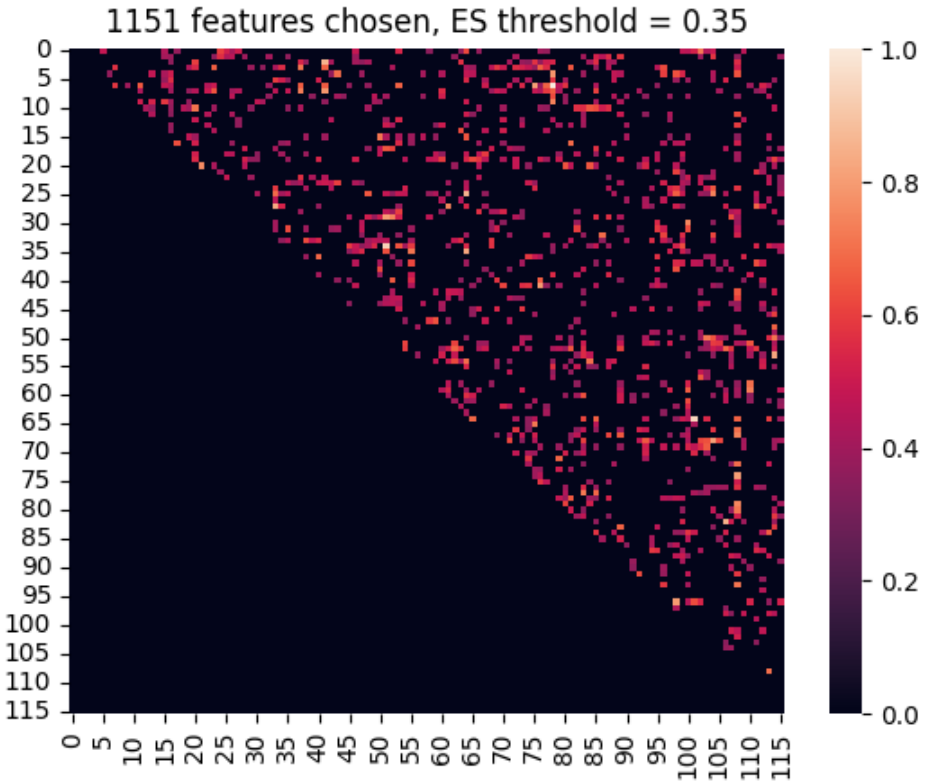


Figure 3.9: The effect size matrix after removing values below the ES threshold. Each cell of the matrix corresponds to an edge in the brain atlas.

The chosen edges were then used as features in place of the input brain networks, and a Probabilistic Neural Network (PNN) was trained for the classification task.

This chapter covered the details of the studied approaches to ASD classification on the ABIDE dataset. The following chapter will summarize the experiments conducted in this study.

Chapter 4

Experiments

This chapter will summarize the history of the work and experiments undertaken in this study, including paths that became infeasible. It will then present the results of the experiments. It should be noted that the entire history of the experimentation phase can be tracked by observing the git history of this thesis's repository, which can be found in Appendix A.

4.1 Replication

The study began with an attempt to replicate the results of Lanciano *et al.* using the replication package provided in their paper, "Explainable Classification of Brain Networks via Contrast Subgraphs" [1]. Links to their repository and the repository for this study can be found in Appendix A. However, unfortunately, no code was provided (by the time of writing) for extracting features using contrast subgraphs or for evaluating the resulting model as discussed in Section 5.1 of their paper. The only code present was for finding a CS given a group of brain networks. Furthermore, the code that was present for extracting a contrast subgraph, based on the work of Cadena *et al.* [52], was not conducive to large-scale experiments. The impact of having insufficient artifacts for reproducibility will be discussed in Chapter 5.

With this knowledge, a rewrite and optimization of the existing code such that it could be run repeatedly for the experiments began. After inspecting the existing code, the most notable inefficiency was the writing and reading of multiple files and the use of subprocess calls to multiple scripts written in Python and Matlab. All code written in Matlab (primarily the SDP solver) was replaced with a Python im-

plementation using the CVXPY library[1]. The solvers were tested to ensure that the same inputs produced the same outputs, so as to prevent any logical changes from this optimization. In the "experiments" directory of this study's repository, two files named matlab_sdp.npy and python_sdp.npy can be found. These files contain example outputs of the Matlab and Python SDP solvers respectively for the same given input. It was found that the outputted values differed only on the order of 10^{-6} to 10^{-4}, which is likely due to a difference in precision between the libraries used.

Rather than passing data through the writing and reading of files and the use of subprocess calls to run new scripts, the scripts were modified to directly accept such data as arguments. The code was then updated from Python 2.7 to 3.8 and unnecessary libraries were removed. The last step of the local search algorithm described in Section 3.1.3, namely the assessment of the complement node-set, appeared to be redundant, and during the translation of the code, this step was accidentally omitted. A test was later carried out in which 5 difference networks were each used with 1000 random node sets in the local search algorithm. Over the 5000 iterations of the algorithm, the complement of the improved node-set was never superior, hence it was very unlikely it could have affected the results of the replication.

Some of the original components from the provided code were reused, but many modifications were needed. The modifications were tested to ensure logical equivalency. After these changes, a logically equivalent implementation of the original code was obtained with increased computational efficiency.

However, as discussed in Chapter 3, finding CSs is a small fraction of the work needed to reproduce the experiments described by Lanciano *et al.* The CSs must be used to translate a brain network into a vector or point. Because this translation was not present in their repository, it was necessary to write from scratch the components for the embedding such as inducing subgraphs, counting overlap between CSs and new brain networks, and calculating the L_1 norm. It was also necessary to write the code for running the cross-validation experiments and using the classical models from scikit-learn[2], but this will be discussed further in Section 4.2.

The replication was attempted using the best, chosen alpha parameters as reported by their paper. Unfortunately, the resulting contrast subgraphs were empty due to the alpha values being too large, and instead, a tuning method was used for determining an adequate alpha value. However, later in the experimentation phase, the authors

[1]This library can be found at `https://www.cvxpy.org/`.

[2]Found at `https://scikit-learn.org/stable/index.html`

released a clarification that the reported values were percentiles, not alpha values. These percentiles correspond to the ratio of edges that are shifted to be considered detrimental to the CS objective (e.g. a percentile value of 70 indicates that the edge weights should be decreased such that only the heaviest 30% of the edges retain a positive edge weight).

After multiple iterations of experimentation and identifying and fixing bugs in the code, a replication experiment was conducted as closely to what Lanciano *et al.* described as possible. This included using the percentile values reported as being the best parameters in their paper and performing 5-fold cross-validation. After translating the brain networks using the CS approach, the StandardScaler class provided by the scikit-learn library was applied to normalize the data points in each dimension. Finally, Lanciano *et al.* allude to using an SVM-based classifier for predicting the classes of the brain networks, therefore, the classic SVC model from scikit-learn's SVM module was used with an RBF kernel. This same model was used for all approaches in the experiments to provide a fair comparison. The results can be seen in Table 4.2 in Section 4.4 as well as in the experiments/replication directory in this thesis's repository (this also contains the exact parameters used along with other metrics). Unfortunately, the accuracies obtained were not as high as reported by Lanciano *et al.*

Based on the discrepancy between the results and the state of the original repository, which used all the brain files in a directory for finding a contrast subgraph, it was hypothesized that Lanciano *et al.* may have found contrast subgraphs using all of the brain graphs before running the 5 fold cross-validation. A small experiment was performed to recreate this possible scenario. Alpha values were obtained from the entire group of brain networks within each category based on the best percentiles reported by Lanciano *et al.* (see percentile_alphas.ipynb in this thesis's repository). The alpha values were then fed into the original code in Lanciano *et al.*'s repository, which uses all of the brain networks of each category to find contrast subgraphs. The contrast subgraphs obtained were then used to translate brain networks into 2D vectors in 5-fold cross-validation. With little tuning, 79% accuracy was easily obtained for the children dataset. However, not much additional time was spent on this experiment, nor are results reported in this thesis as this is completely invalid. The approach involves the test data when constructing the prediction model. This experiment was only conducted to try to explain Lanciano *et al.*'s high-accuracy reporting.

In their paper, Lanciano *et al.* describe a nested cross-validation technique in

which parameters were optimized. To be completely fair to their approach, it was necessary to develop a module for performing such experiments. The following section will describe the work done as it applies to the experiments for every approach.

4.2 Evaluation Framework

To ensure all approaches were evaluated fairly, a common framework was developed to run the experiments. This framework went through various revisions and iterations as did the experimentation phase. At one point, the pipeline module provided by scikit-learn was considered, but due to a lack of documentation regarding observed behaviours (specifically around instantiating custom transformer classes), it was decided that the best approach would be to recreate a simplified version of the module for this study.

The pipeline module developed for this study works very similarly to the module provided by scikit-learn, but it instantiates all of the steps of the pipeline once with a given set of parameters. It can also plot transformed data points in certain scenarios. 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 would receive the brain networks as input and output the feature vector specific to each approach. As mentioned in Section 4.1, 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 classifier from scikit-learn using an RBF kernel, which is a very common classifier in ML.

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 hyperparameter optimization as follows [1]:

We randomly split the data into 80/20 training/test subsets. Using the

> *training set and 5-fold cross-validation we select the best hyperparameters for the classifier. We then apply the best classifier to the test set. We repeat this process 5 times for each method.*

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[3]. 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. See Chapter 5 for a discussion on why accuracy was chosen for this as well as emphasized throughout this study.

The scikit-learn documentation has a discussion and demonstration of the differences between nested cross-validation and non-nested cross-validation[4]. It also references the work of Cawley and Talbot [54] concerning the risk of over-fitting models to a specific dataset when using only standard cross-validation. Figure 4.1 illustrates the approach taken in this study. Moreover, a standard grid search with non-nested cross-validation was used for comparison.

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:

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

[4]See `https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html`

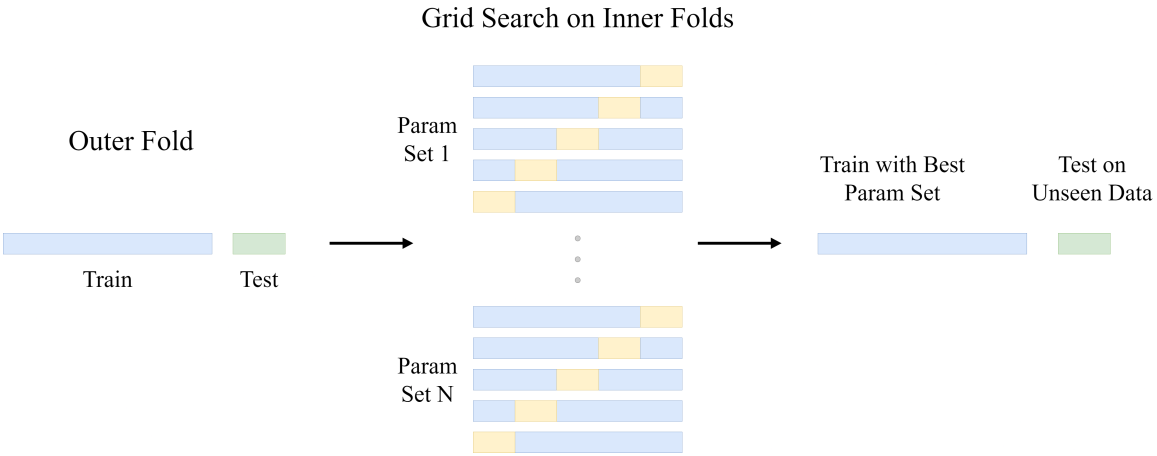


Figure 4.1: A single fold of the nested cross-validation scheme used in this study. This sequence is repeated for all of the outer folds of the data (in this study, 5 folds were used for the outer and inner folds). Note that only the train data is used during the grid search. For each combination of hyperparameters, cross-validation is used on the train data. The set of parameters achieving the highest average accuracy is used to train the model with all of the train data before predicting the test set.

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.

For the effect size thresholding method, the thresholded effect size matrix (as in Figure 3.9) was plotted for each fold.

4.3 Procuring Data

The early experimentation phase used the thresholded brain networks provided in the repository of Lanciano *et al.*'s work [1]. However, as previously mentioned, many approaches have achieved better results using raw correlation matrices. Therefore, a script was created to download the appropriately preprocessed BOLD time series files from the PCP [33]. Here "appropriately preprocessed" simply means "in accordance with what Lanciano *et al.* claim to have used". The DPARSF[5] pipeline was used with band-pass filtering and global signal regression, and the version of the AAL[6]

[5]See `http://preprocessed-connectomes-project.org/abide/dparsf.html`

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

brain atlas provided by the PCP was chosen.

The PCP provides a file named "Phenotypic_V1_0b_preprocessed1.csv"[7], which can also be found in the data/ABIDE directory of this study's repository. This file provides information about each subject such as their subject ID, filename, and various phenotypic traits. Some of the subjects did not have valid file names, but it was possible to reconstruct most of them using values from other columns and some manual work. Out of the 1112 subjects claimed to be provided by the ABIDE dataset [2], 1102 were able to be downloaded.

As the files were downloaded, they were sorted into the various categories outlined by Lanciano *et al.* in their study (e.g. male and children) according to the criteria described in their paper. The categories were not mutually exclusive (e.g. a male child would be found in both the male and children categories). Additionally, the categories did not encompass all subjects. Specifically, females that were not instructed to close their eyes during the scans and were not classified as children or adolescents were excluded from Lanciano *et al.*'s study. This category was labelled as "other" in this study (as seen in Table 4.1), though it was too small of a group to warrant any reliable experiments, so it was not used independently. Instead, a new category named "all" was created using a script to determine the unique subjects across categories. This included the "other" category and is in line with many other studies that use the ABIDE dataset as a whole.

To closely follow the procedure described by Lanciano *et al.*, any BOLD time series with missing or null values were removed before converting from time series to correlation matrices. However, as made clear by Table 4.1, there were many inconsistencies between the number of subjects retrieved in this study and theirs. For most categories, this study retrieved more subjects. This may be due to the effort to reconstruct filenames where they were not present in the phenotypic CSV file provided by the PCP (and originating from the ABIDE dataset). It may also arise from a slightly different method of filtering out subjects, though if there is a difference in the methods used, it was not described by Lanciano *et al.*. However, peculiarly, this study retrieved fewer files for subjects in the children category, as well as subjects with ASD in the adolescents category. The cause for this discrepancy is uncertain.

[7]See `http://preprocessed-connectomes-project.org/abide/download.html`

Table 4.1: Subject counts for the present study by category, file type, and class. The first four categories are defined exactly as Lanciano *et al.* define them in their paper [1]. 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 obtained from the BOLD time series in this study. The "Downloaded Timeseries" column corresponds to the BOLD time series downloaded from the PCP.

Category	Lanciano (thresholded)		Raw Correlation		Downloaded Timeseries	
	ASD	TD	ASD	TD	ASD	TD
children	49	52	40	39	41	39
adolescents	116	121	114	122	121	125
eyesclosed	136	158	141	165	164	183
male	420	418	443	455	467	472
other	0	0	27	49	27	49
all (unique)	457	462	504	551	531	571

4.4 Results

After many iterations of rigorous research and experimentation, results were obtained that were deemed trustworthy. 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 4.2 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 4.2: Replication results. This is modelled after Table 2 in Lanciano *et al.*'s paper [1] 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. As described in Chapter 3, there were multiple variations and approaches considered as well as implemented. 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.

- CSP2-SDP-N1 - The recreation of CSP2 from Lanciano *et al.*'s work.
- CSP1-QP-N3 - The CSP1 approach with modifications as discussed in Section 3.1.3. QP specifies the solver used, and N3 specifies the number of contrast subgraphs used.
- CSP2-QP-N3 - The CSP2 approach with modifications as discussed in Section 3.1.3.
- DE - The Discriminative Edges approach as described in Section 3.2.
- Iidaka - The effect thresholding approach as described in Section 3.3.

The approaches for problem 1 of the contrast subgraph technique (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 problem 2 of the contrast subgraph technique (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 the 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. Furthermore, for the time-consuming approaches, manual testing on smaller parameter grids was conducted until an appropriately sized parameter grid could be used that took a reasonable amount of time to run, but gave the approaches a fair chance at performing well. It was also ensured that the parameters listed in their paper were included in the grid search.

Figures 4.2 to 4.5 provide average prediction accuracies (with standard deviations) as well as average runtimes for training the pipeline for each approach. As previously

mentioned, nested cross-validation and non-nested cross-validation experiments were performed, both using grid searches for hyperparameter selection.

Several observations can immediately be made after inspecting these charts:

1. None of the accuracies are particularly impressive 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.

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.

An in-depth discussion of the results shown here can be found in Chapter 5. 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.

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

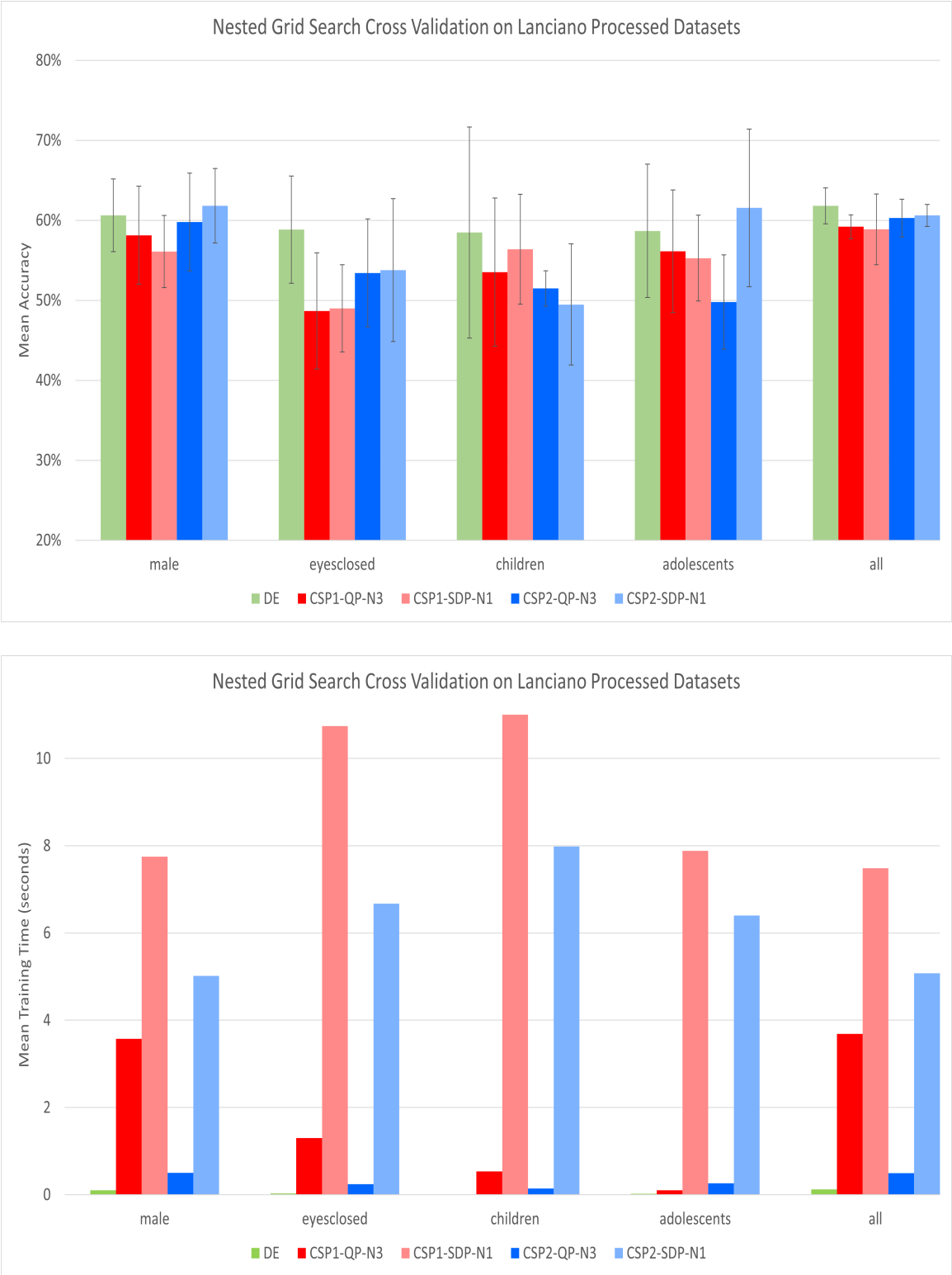
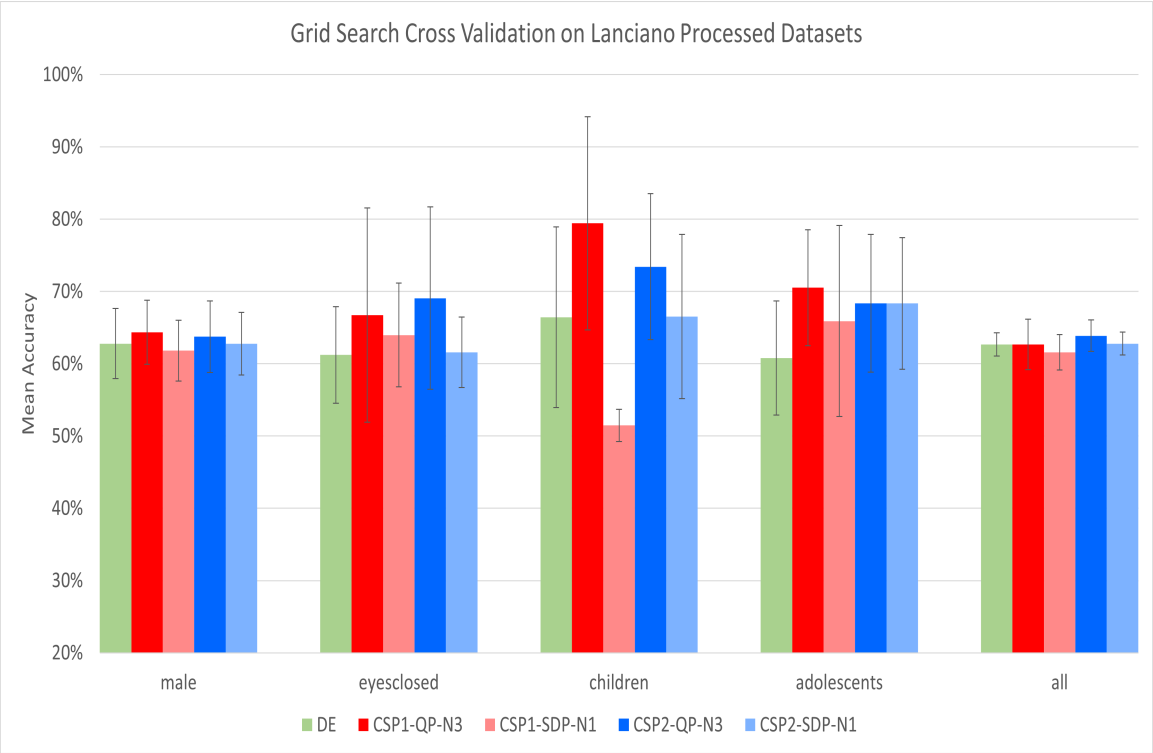


Figure 4.2: Thresholded-NestedCV Results.



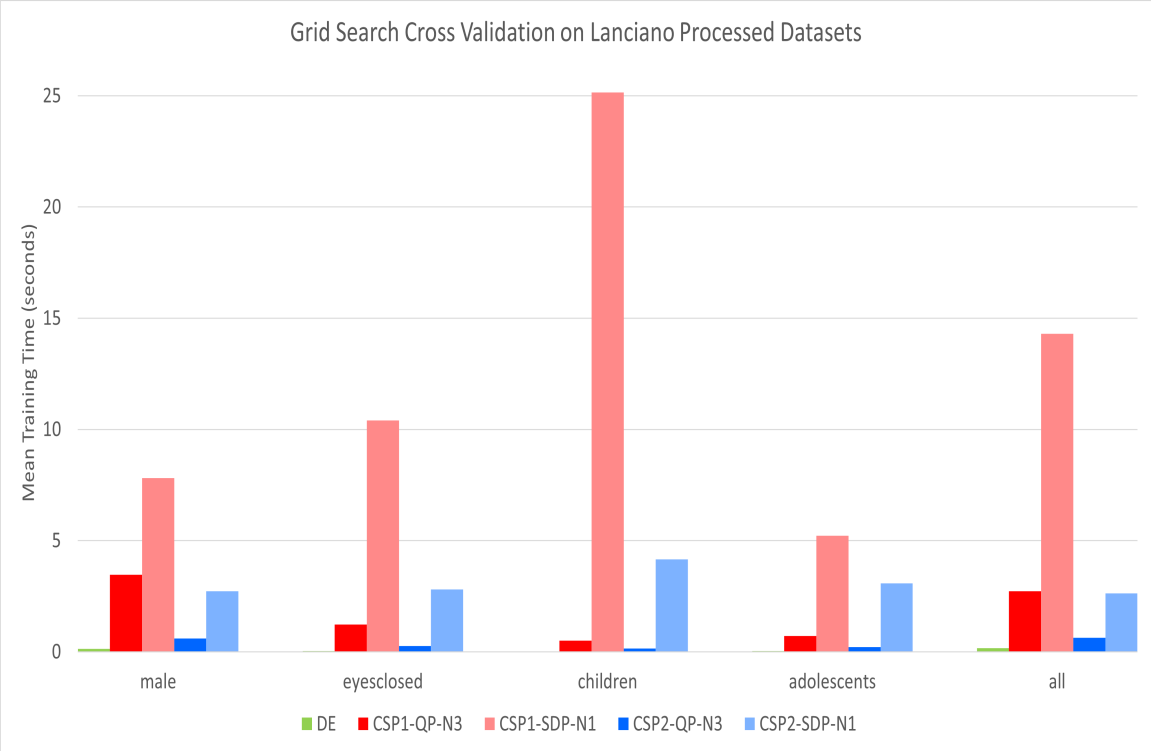


Figure 4.3: Thresholded-CV Results.

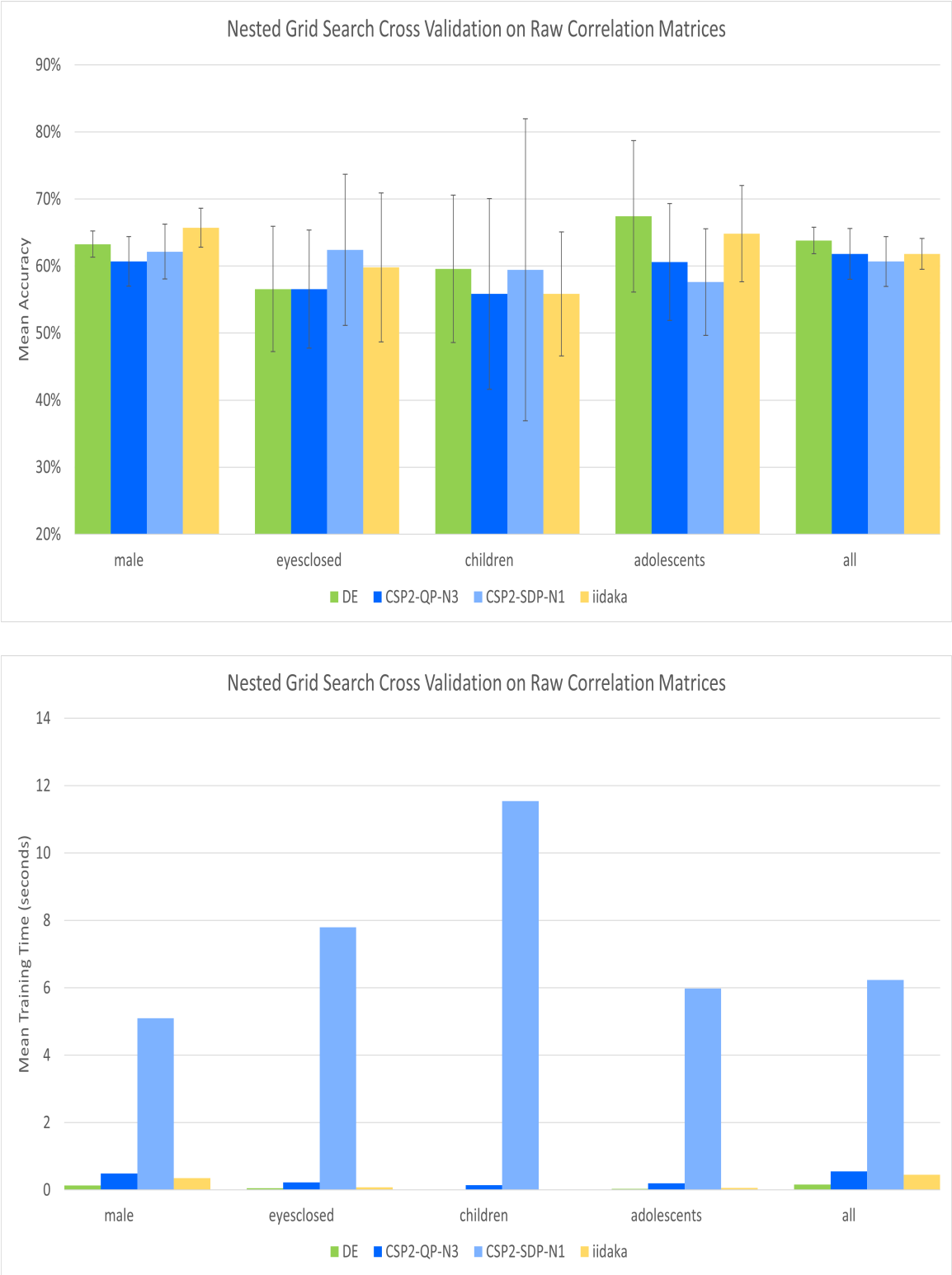
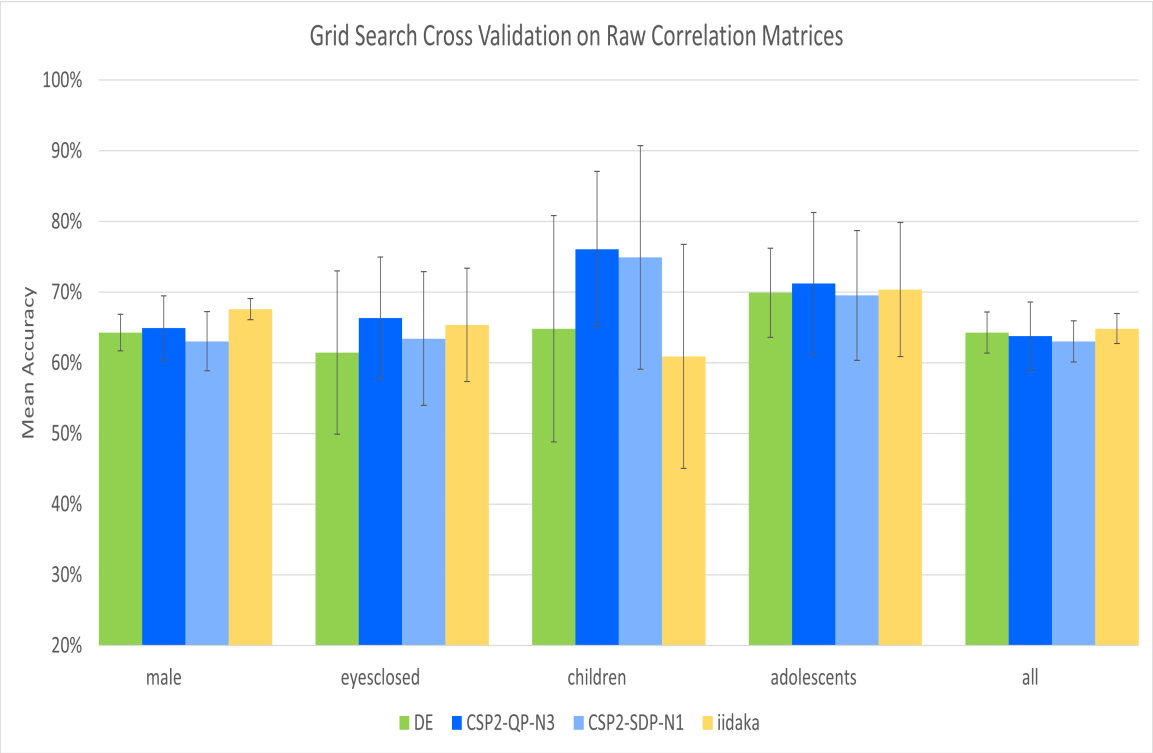


Figure 4.4: Raw-NestedCV Results.



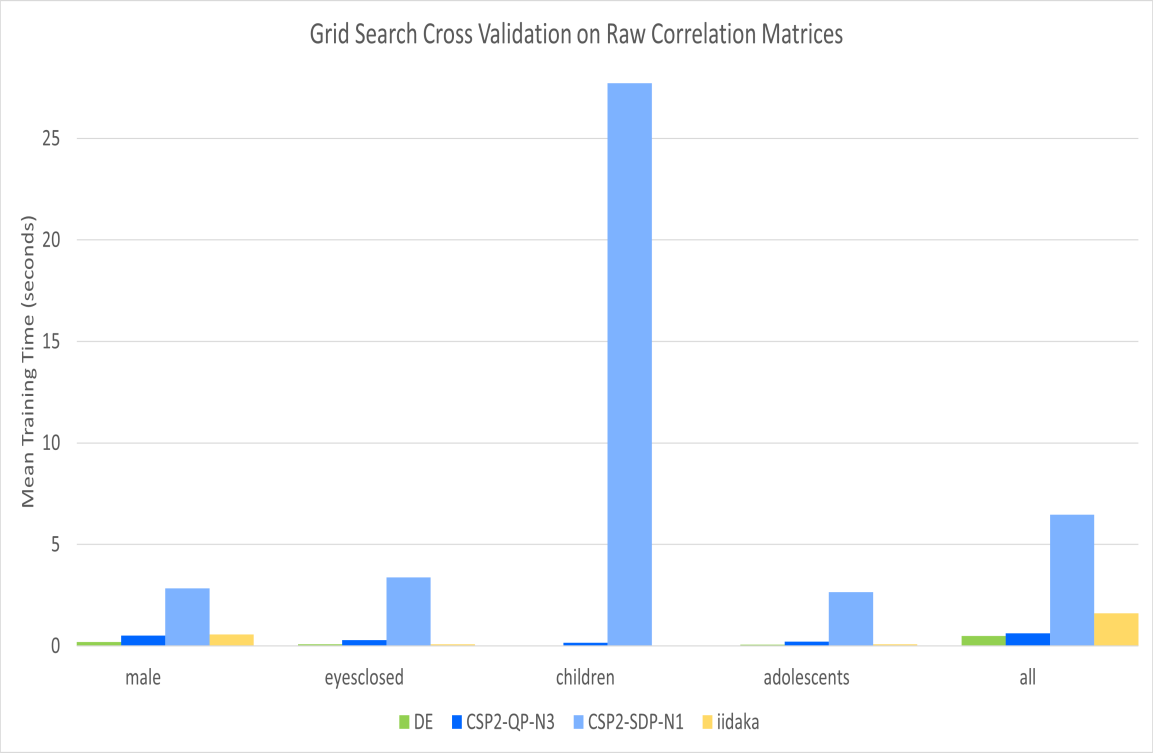


Figure 4.5: Raw-CV Results.

the 50 folds. The adjacency matrices were then thresholded to show only the most discriminative edges. These connections were visualized with the BrainNet Viewer [55] as seen in Figures 4.6 and 4.7.

4.5 Failed Experiments

In research, it is useful to know the ideas and experiments that were explored but did not yield favourable results. This section will briefly describe some such failures and explored paths.

The graphs used in this study have a special attribute called Node Identity Awareness (NIA). This means that the nodes of the graphs (in this case, ROIs of a brain), are the same between graphs. Hence, there is a fixed number of nodes and edges in the brain networks. Therefore, an attempt was made to simply pass the unfiltered brain networks and correlation matrices into simple SVM and DNN classification models using every connection in the brain as features. The results were not competitive, which may be attributed to the amount of noise contained in the networks. Despite this, it may have been useful to run experiments with this approach to serve as a baseline to compare against the other approaches.

Earlier in the experimentation phase of the contrast subgraph approach, before using percentiles to determine the values of the α hyper-parameter, a tuning phase was implemented to find an optimal value of α. Initially, the tuning phase focused on maximizing the accuracy obtained with certain values of α, but because of the poor performance, a new attribute was considered for maximization. Recall Section 2.3 in which separability was proposed as a major aspect to consider when selecting features. In an attempt to find a value that measured the extent to which data points were separated between classes, the Distance-based Separability Index (DSI) was discovered [56]. Unfortunately, maximizing the separability (as measured by the DSI) of the training data did not achieve adequate accuracies, possibly indicating the CS approach lacked extensibility.

As an extension to the concepts in the problem 1 formulation for the CS approach, an idea was considered for counting the overlap of higher-order graph structures. An initial attempt was made to count triangle overlap before proceeding to a wider variety of graphlets involving more edges. The hope was that perhaps certain patterns of connections were more or less common in the brains of people with ASD and that such patterns could be identified for classification. Unfortunately, not much success

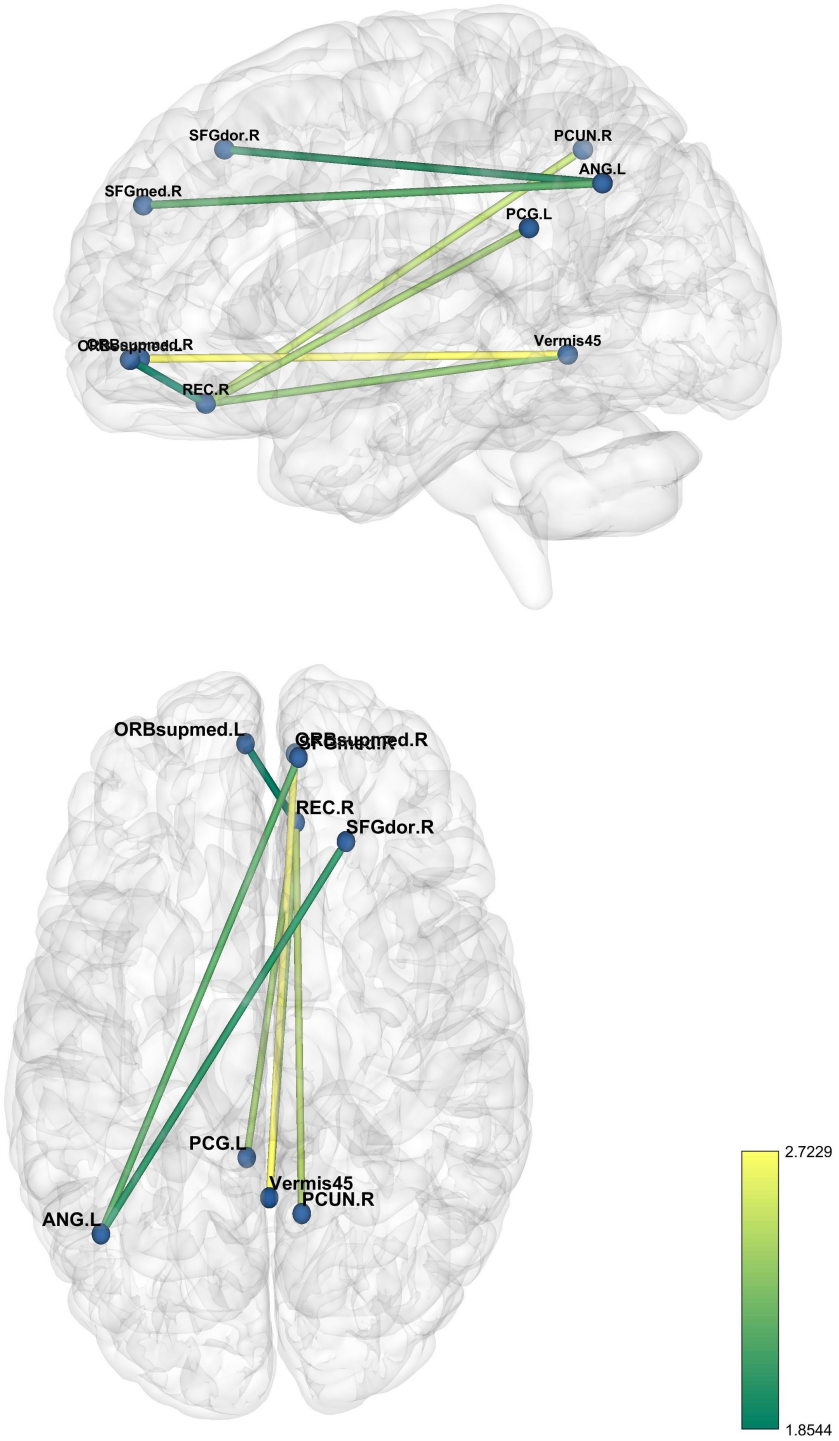


Figure 4.6: Connections that were more highly correlated in the brains of typically developed individuals. Top: Sagittal View. Bottom: Axial View. The range of edge weights is indicated by the scale in the lower right. Edge weights represent the respective sum of the difference network edges when each edge was selected during the 50 test folds. Note the ROI with the highest degree is the right Gyrus Rectus.

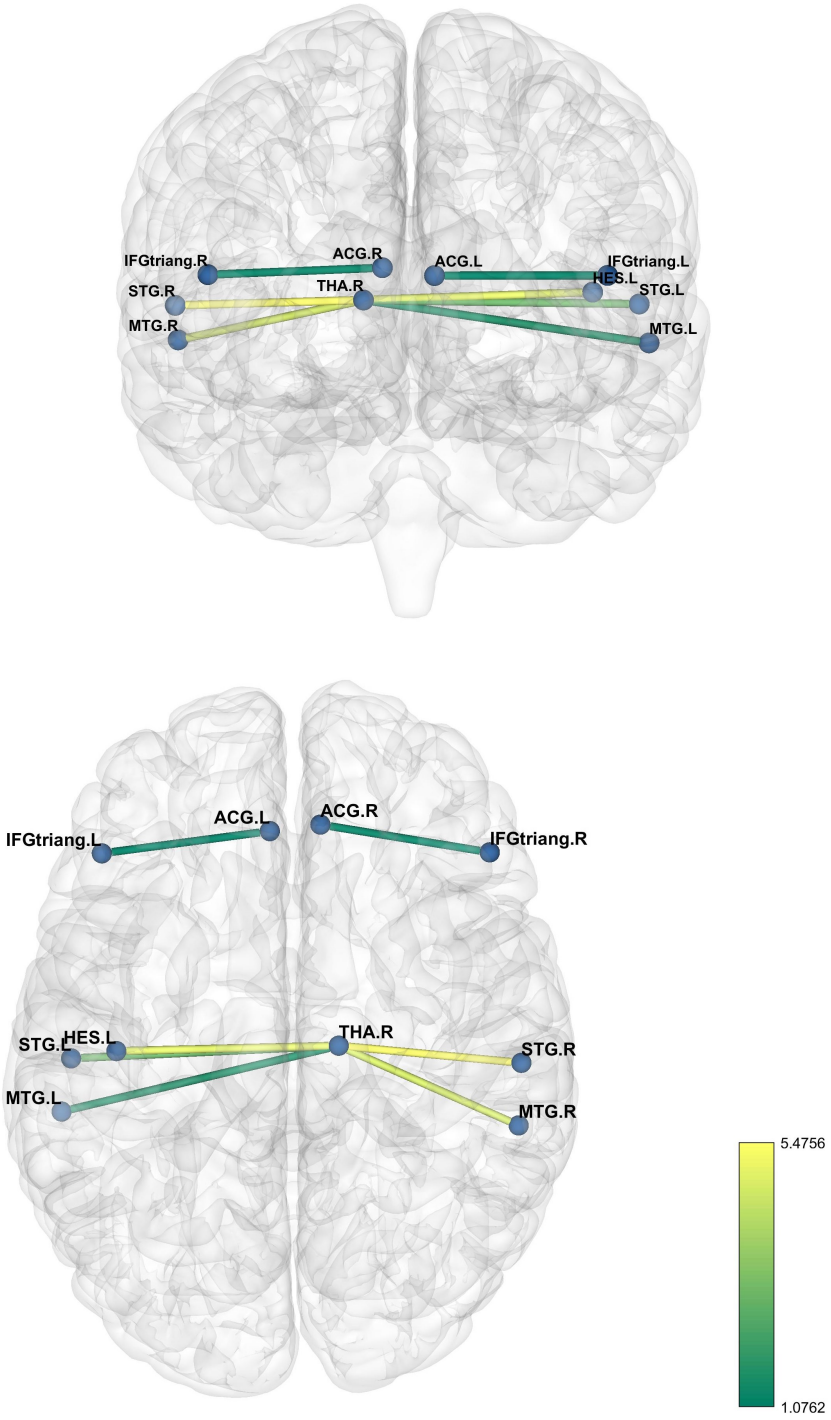


Figure 4.7: Connections that were more highly correlated in the brains of individuals with ASD. Top: Coronal View. Bottom: Axial View. The range of edge weights is indicated by the scale in the lower right. Edge weights represent the absolute value of the respective sum of the difference network edges when each edge was selected during the 50 test folds. Note the ROI with the highest degree is the right Thalamus.

arose from the experimentation with triangle overlap, and the computation time was significant (and would only increase with higher-order graphlets).

An attempt was made to recreate the thresholding of the correlation matrices as done by Lanciano *et al.* This was moderately successful, as the example brain networks examined seemed to be equivalent except for one or two edges (it is unclear what the differences would have been in thresholding the values between the two studies). With the code in place to perform the thresholding at various percentiles, new datasets were generated to mimic those provided by Lanciano *et al.* but at the $70^{th}, 75^{th}, 85^{th},$ and 90^{th} percentiles. Various experiments were conducted using these new datasets, but the results were not as favourable as those thresholded at the 80^{th} percentile. For the sake of having a more direct comparison to their paper, the brain networks provided by Lanciano *et al.* were used in the final experiments rather than those generated in this study.

Various other experiments and paths were explored that were not deemed noteworthy enough to be mentioned in this section.

Chapter 5

Discussion

Various observations were made throughout this study. This chapter will discuss and analyze these observations as well as discuss the limitations of the study, the importance of reproducibility, and ideas for future work in the hopes that the insights gained here can be useful in this area of research.

5.1 Analysis

Section 4.4 made some basic observations about the results of the experiments. The first observation is the most obvious, but it is not clear why this is the case. As discussed previously, ASD is a complex disorder and it may manifest in various ways. To add to that, the brain itself is a complex organ that is not fully understood, and the current methods and technologies for studying the brain have much room to develop. There are likely a plethora of confounding factors that make this problem difficult. For example, one study found that the median percentage of autism cases with co-occurring intellectual disabilities was 33% [10]. The brain networks of those individuals would likely appear noticeably different than the others despite all of the individuals possessing ASD. Considering ASD is also a spectrum, and it can affect individuals to very different degrees, it might not make sense to view this as a binary classification problem as most research in this area does.

This study also uses fMRI data that was obtained from multiple different scanning sites that did not follow a common study procedure. Multi-site classification has been observed to be more difficult than single-site classification [44]. A noticeable difference in the procedures of the sites contributing to the ABIDE dataset is the length of the

BOLD time series. It is very possible that the length of the scans could affect the value and stability of the derived correlation coefficients.

The second observation will be discussed in Section 5.1.1 and relates to **RQ1**.

The third observation relates to **RQ2** and may indicate that there is a limiting factor to the approaches implemented in this study. This may be due to the limitations of the dataset or the classifier model used, or perhaps the approaches are too similar, and more avenues should be explored concerning data preprocessing and feature selection. This will be discussed further in Section 5.2.

The fourth observation appears to answer **RQ3**. However, there are a few factors to consider. Firstly, although the difference appears noticeable when reviewing the results, there may not be a statistical difference. Additionally, the approaches used for the different types of inputs are different for the most part. Even in the case of DE, which takes both input types, its algorithm changes based on the input type, so the comparison may not be fair. The potential advantage of thresholding the correlation matrices is to reduce noise, but these results, along with the empirical results obtained while experimenting with different threshold values as mentioned in Section 4.5, seem to indicate that the common approach of utilizing the raw correlation matrices may be superior as it includes more information.

The fifth observation is simply that, of the approaches studied, DE is the most computationally efficient, which provides an answer for **RQ2** in terms of performance. This means it could more easily integrate the information from multiple atlases, especially those of higher resolution. The model made by Epalle *et al.* mentioned in Section 2.4.2, on the other hand, takes roughly a week to train. It might be fine to have a week-long training time if the classification model were to only be trained once, or very infrequently, but by using DE new subjects could be added to the model quickly and regularly.

The sixth observation will be discussed further in Section 5.1.4.

5.1.1 Discrepancies in Results

The recreated effect size thresholding approach did not perform to the degree that Iidaka claimed, but it should be specified that there were differences in the experimental setup and methods. One major difference comes from the fact that an SVM classifier was used rather than a PNN. Another main difference is that a different version of the AAL atlas appears to have been used in their study (this will be discussed

in Section 5.1.2). Moreover, Iidaka used leave-one-out cross-validation, whereas this study used 5-fold cross-validation. Iidaka also only experimented on individuals under 20, though this study also considered subsets of the ABIDE dataset.

Regardless of these differences, it still does not account for a nearly 30% difference in accuracy. Woo *et al.* discuss the observed biases in studies that perform analysis procedures on the entire dataset before splitting the data into training and testing sets [16]. It is unclear what Iidaka did based on the description of their methods, but they do not explicitly mention recreating the effect size matrix for each fold of the leave-one-out cross-validation, which would have been considerably more effort than creating the effect size matrix for the entire dataset once. This represents a possible leak of information into the model which may have resulted in an optimistic bias.

Similarly, this hypothesis may help explain the discrepancy between this study's results and the results reported by Lanciano *et al.* A small experiment was conducted based on this hypothesis, as discussed in Section 4.1, and it found that higher accuracies were easily achievable when contrast subgraphs were found using the whole dataset before running the cross-validation. The code repository provided by Lanciano *et al.* was insufficient to conduct the experiments in a valid way based on their description of the experiments.

These uncertainties highlight the need for greater reproducibility, which will be discussed in Section 5.3. The repository containing the code and data to replicate this study's findings and verify its methods can be found in Appendix A.

A peculiarity of the results can be spotted when analyzing the graphs of the training times. The SDP approaches took a significantly longer time to train than the other approaches. However, notice how the children category consistently had longer training times than the other categories. This is quite odd as the children category is the smallest category among them, and the trend even continues, for the most part, with categories including more subjects taking less time to train. This behaviour is only observed in the SDP-based approaches, as the other training times behave as one would expect, with larger categories taking more time to train. Perhaps there were issues with caching during the execution of the SDP approaches, but this is purely speculative.

5.1.2 Brain Atlas

A broken link exists on the PCP's website[1] concerning the AAL atlas used. It was noticed that the work of Iidaka [28] and Epalle *et al.* [45] both claimed the use of the AAL atlas, but they obtained 90 ROIs compared to the 116 ROIs obtained in this study and by Lanciano *et al.* [1]. It is therefore uncertain what exact version[2] of the AAL atlas was last used and released by the PCP [57, 58, 59].

It is also worth noting that the AAL atlas may not be superior to other atlases. Subah *et al.* observed that the AAL atlas produced inconsistent results and had the lowest sensitivity compared to the other atlases used in their study [41].

5.1.3 Correlation Coefficients

When creating brain networks from fMRI data, it appears that nearly all of the studies in this area of research compute the Pearson correlation coefficients between the ROIs of the brain. However, this might not yield the most information-rich data. When creating the thresholded brain networks mentioned in Section 4.5, it was noticed for a couple of subjects' brain networks that the 80^{th} percentile of the correlation values was around 0.2. This is often considered an indication of a weak correlation, and indeed, as seen in Figure 2.2, there are not many values above 0.7, which is generally considered a strong correlation [60]. There could be various reasons for this observation. Perhaps this indicates that the Pearson correlation of ROI time series is not the best way to measure functional connectivity, or perhaps this could simply point to errors in data retrieval or preprocessing. It is also possible, though unlikely, that the inspected correlation matrices were abnormalities. Whatever the reason, this warrants further investigation.

Additionally, most approaches ignore the possible useful information from negatively correlated BOLD time series. It could be that one region may activate in response to another region, but with a slight delay, so the time series is shifted in time. Figure 2.1 showed the time series of three ROIs. The green time series has a negative correlation with the others even though it may actually fluctuate in response to the activity of those ROIs. Current approaches treat this relationship as being equivalent to having no correlation at all, but it might be worth distinguishing

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

[2] For available versions, visit `https://www.gin.cnrs.fr/en/tools/aal/`

between negative correlations and the lack thereof to obtain more useful information. It should be noted, however, that the GSR processing step is thought to introduce negative correlations that may not exist in reality [32].

Furthermore, the correlation values may not give the whole picture. For example, it may be that one individual has a similar correlation between two ROIs as another individual, but perhaps the first individual has far more intense fluctuations of BOLD signals than the second individual [60]. The correlation value does not indicate what the BOLD signal of a region is doing objectively, only its activity relative to other regions. Hence, more interesting differences may exist between the ASD and TD groups in the way these BOLD time series behave.

5.1.4 Issues with Contrast Subgraph Approaches

Note the accuracy values of CSP1-SDP-N1 on the children category in Table 4.2 and Figure 4.3. The experiments in this study used random state seeds wherever possible to ensure reproducible results. Additionally, the parameter grid used for the grid search cross-validation included the parameters used in the smaller replication experiment. Therefore it should be the case that the accuracies of the grid search cross-validation would be at least as high as the replication. After some investigation, it was determined that the random projection step in the SDP solver function (which originates from the code provided by Lanciano *et al.*) may be the cause of the inconsistent results. This could be given a random state seed to achieve consistent results, but this helped to identify just how unstable the SDP method can be, especially in the case of the children category, which has very few subjects.

As mentioned previously, many iterations of experimentation were conducted. Occasional bugs and issues in the code were found in each iteration. One such bug was an off-by-one error for the SDP implementation when returning the solution vector. Despite having such a major issue, the technique achieved similar results before and after detecting and fixing the bug. It appears that the local search step was sufficient to amend the solution vector so the results were not much different (which is what led to this error going undetected for some time). This highlights the possible superfluousness of the SDP solver, given that it accounts for a majority of the computation time of the approach.

Similarly, an issue was detected with the first implementation of the QP solver and after solving this issue, no noticeable performance increase was noticed. This

perhaps indicates that these approximations of the solution to the GOQC problem are insufficient and unnecessary.

Another main issue with the CS approaches has already been mentioned in Section 3.2. The motivation for developing the DE approach was to move away from finding a set of ROIs that induced a contrastive subgraph and instead determine a set of connections between ROIs that were discriminative between the two classes. When using a set of ROIs, connections are included that do not give much information about the class of the brain network, yet they are given equal weight with the more important edges when determining feature values.

These findings seem to indicate that finding a dense subgraph overcomplicates the problem. There is no need to identify a connected and complete subgraph when individual connections are the ones that provide the useful information. Rather than approximating an NP-complete problem (i.e. GOQC as given by Cadena *et al.*), the most discriminative edges can be linearly determined and analyzed.

5.1.5 Focus on Accuracy

Though confusion matrices and other metrics like f1-scores are recorded in the provided repository for this study's results, the accuracy of the approaches was emphasized. Accuracy is often used when the classes are balanced (which is the case for the ABIDE dataset) and there is no major downside to predicting false negatives. There is certainly a downside to predicting false negatives in this context, and therefore a high accuracy is not sufficient for this problem. However, achieving a high accuracy is certainly necessary for such models to be trusted in practice. Given that no technique has been able to attain a suitably high accuracy on the ABIDE dataset, it makes sense to first make breakthroughs in this area before focusing on other metrics.

5.1.6 Nested Cross Validation vs Cross Validation

Nested cross-validation only makes sense in the context of hyper-parameter tuning, which is an essential step in building a classification model. Grid search was used for parameter tuning in this study, as can be seen in the middle step in Figure 4.1.

When non-nested cross-validation is performed, various hyper-parameters are tested, and in the end, those achieving the best results are chosen. Unfortunately, this leads to an optimistic bias as the model may be over-fitted to the testing data. Nested

cross-validation simulates a more realistic environment, and helps to identify if the model is extensible to new data.

While determining appropriate grids for each approach's parameters, an interesting observation was made that illustrates the key difference between nested cross-validation and non-nested cross-validation. While performing nested cross-validation, two parameter grids were tested, the second grid was a superset of the first grid. However, the round of experiments using the second grid led to lower accuracies because the parameters chosen during the grid search were over-fit to the training data. Though the score was lower, it was more indicative of how the model might have scored when predicting classes for an entirely different dataset that it had never trained with.

This is why the results of the nested cross-validation experiments indicate a noticeable decrease in accuracy compared to the non-nested cross-validation experiments.

5.1.7 Explainability

Figures 4.6 and 4.7 give some insight as to what connections in the brain were found to be most important for discriminating between the classes by the DE algorithm. However, this form of explainability could be applied to just about any approach assuming that the derivation of the features involves choosing certain connections. For example, Lanciano *et al.* present similar diagrams to visualize the contrast subgraphs found. For Iidaka's effect size thresholding approach, one could apply an even higher threshold on the effect size matrix to reduce the number of edges chosen and display the connections corresponding to the most discriminative edges. For approaches that simply feed the whole correlation matrix into a black box classifier, there are tools such as SHAP that can identify the most important edges learned for classification.

However, explainability also comes partly from the way features are derived. With the DE algorithm, it is clear why certain edges are deemed important, but with a black box classifier, this is less clear. In this study, the features have been either correlation values or values derived from correlation values, but this does not need to be the case. Features can be derived in many different ways, such as with graph analytics and metrics [61].

Explainability also varies with the classification model used. Many classic ML models have straightforward algorithms that can be easily understood by experts in the tools. Decision trees are known for being highly explainable. They provide

classifications that can be understood by non-experts, though more knowledge may be needed to understand how the decision boundaries are calculated.

Explainability without accuracy is insufficient though. Perhaps better than guessing and checking what features of the brain provide good predictive power, it would be useful to work backwards from the classifications of powerful black box classifiers (once a sufficiently accurate classifier is developed) and use graph metrics involving the connections identified as important by explainability tools such as SHAP.

5.2 Limitations

This study, as well as most of the other studies with the same goal, suffers from a lack of appropriate data. Most areas of research that apply machine learning utilize datasets with thousands, millions, or even billions of records depending on the context. Small sample sizes have been observed to lead to unstable classification models, whereas larger samples have been shown to increase classification accuracy along with stability [15]. This may partially explain the lower accuracies obtained in this study, though it does not explain the discrepancies between the accuracies in this study and those reported by others. The stability of the models can also be observed in the standard deviation bars in Figures 4.2, 4.3, 4.4, and 4.5 where categories containing more subjects exhibit smaller standard deviations.

A major goal in this area of research is to provide early diagnoses for infants and toddlers. However, one issue with respect to retrieving fMRI data for younger subjects, is that scans require subjects to remain as still as possible throughout. Because the change in blood flow during neural activity is not immediate (it can take about 5 seconds to reach the maximum flow [30]), the brain must be scanned over a fairly long duration of time to retrieve results that can convey useful information. Moreover, studies on such young subjects require follow-ups to determine what diagnoses the subjects were given, if any. Redcay and Courchesne conducted a study with younger subjects that were sleeping during scans, but with a very small sample size [62]. This leads to questions of whether it is practical to perform enough fMRI scans on babies to propel this area of research to a point where it becomes useful in practice. If it advances enough to become a reliable source of diagnosis, will MRI scanning be accessible enough for it to make an impact on the general population? Perhaps it will, or perhaps other measurement avenues which are more accessible and practical should be considered and studied.

This study was also limited by time and the breadth of experimentation that could be performed. For example, the choice was made to use only SVM classifiers for each approach for the sake of consistency. However, it could be beneficial to apply DL algorithms to the approaches using the derived features. Table 7 of Subah *et al.*'s study seems to indicate a major advantage of using a deep neural network over a classical machine learning algorithm such as SVM in this context [41].

5.3 Reproducibility

A large portion of time was used to replicate the work of Lanciano *et al.* This raised a concern regarding the reproducibility of studies in this area of research and perhaps on a broader scale.

This thesis adopts the terminology proposed by Goodman *et al.* as recommended by Hans in a recent article discussing the confusion surrounding the terms reproducibility, replicability, and repeatability [63, 64]. This thesis proposes that every study in this domain should describe its methods well enough such that it has *results reproducibility*. That is, one can re-implement the necessary code to obtain the reported results simply by reading the study and using the provided artifacts such as the data. It should also be straightforward to provide *methods reproducibility*, which is closely related to results reproducibility in the context of the computational sciences [64]. It implies that one can simply re-run the provided code to obtain the same results reported by the original author.

Unfortunately, the work of Lanciano *et al.* possessed neither results nor methods reproducibility based on the effort made to replicate their work in this study. This study also attempted to employ the strategies described by Iidaka using a different experimental setup and different subsets of the ABIDE dataset. Though the results did not agree with the claims of Iidaka, it was not a close enough replication to conclude their claims were erroneous. The replication of Lanciano *et al.*'s work was, however, as similar as could reasonably be made based on their descriptions.

Some examples of inconsistencies observed in their work that led to difficulties in the replication are as follows:

- They claimed that the time series for each subject contained 145 units of time, whereas the ABIDE dataset contains time series of variable scanning durations.
- The code in the provided repository used different projection and local search

algorithms than those indicated by Algorithms 3 and 4 in their paper and bugs were identified in the functions that more closely resembled the given pseudocode.

- They claim that the function $f_\alpha(S)$ in Algorithm 4 corresponds to the edge-surplus definition given by Tsourakakis *et al.*, whereas it corresponds to Cadena *et al.*'s generalization of the function to weighted graphs.

It appears that Lanciano *et al.* repurposed code directly provided by Cadena *et al.* Perhaps these issues originated with Cadena *et al.*'s work, but the algorithms should have been validated by Lanciano *et al.* during their study. Ultimately, based on the work done in this study, the answer to **RQ1** is no.

Without reproducibility, it will be impossible to integrate ML classification techniques into medical practice, which is the goal of this research. It has been observed that a cultural change in the computational sciences may be needed to encourage and incentivize studies that pay attention to making their work reproducible [65]. Some unique challenges will be faced with respect to assessing reproducibility as ML technologies are adopted into clinical sciences. One of the issues noted by Beam *et al.* is the presence of many hidden parameters in ML models that can change from software to software or even between versions of the same software [66]. To address this issue, random seeds should be set wherever possible to remove variability in results as done in this study.

Many tools exist for improving the accessibility and reusability of code repositories. For example, Poetry[3] is a Python package and dependency manager used for this study's code repository which makes it very simple to install the necessary dependencies of the code into a virtual environment. This study provides a code repository with datasets and documentation aimed to aid future researchers in understanding the methods used and to invite inspection or improvement.

5.4 Suggestions for Future Work

Numerous ideas were considered during this study, but many of them were out of scope due to either a lack of time or expertise. The knowledge gained during this study may be beneficial to future researchers and hopefully, the suggestions made here will help to advance this area of work.

[3]https://python-poetry.org/docs/

This study used the ABIDE I dataset, but the ABIDE initiative released a second dataset in 2017: ABIDE II [67]. Unfortunately, significantly fewer studies have employed this dataset compared to ABIDE I. The difference in their adoption is partly due to the time difference between their releases, but it is also likely because preprocessing of the dataset has not been openly shared with the research community as with ABIDE I by the PCP. The PCP helped to inform researchers of available preprocessing tools and detailed the steps and parameters they used to process the first ABIDE dataset, but many researchers may not have sufficient background knowledge to ensure the tools are used properly or, at the very least, performing the preprocessing themselves represents a set back in how much time can be used to experiment on the resulting data. As discussed in Section 2.1.1, the PCP's effort on ABIDE I has been incredibly useful in making the data more accessible to the broader research community, and it has reduced redundancy while creating common data for comparisons. An immediately useful step in this area of research would be to perform the preprocessing for the ABIDE II dataset and host the resulting data publicly as the PCP did for ABIDE I.

PCP also provided metrics for assessing the quality of the preprocessed ABIDE data, but not many studies with the same goal as this study have utilized this information. This could be a valuable source of information for filtering the dataset to reduce noise and potentially the number of outliers.

When a successful classification technique is identified, such as the proposed method by Subah *et al.* [41], it should be tested with various preprocessing steps, parameters, and atlases to ensure the technique is picking up on a biological feature rather than a feature of the preprocessing. The PCP provides numerous datasets for each processing pipeline, brain atlas, and with and without GSR or with and without band-pass filtering. Hence, there are many different datasets available to compare a given technique across.

Such studies should be made reproducible, and others should recreate the studies to validate the results. Bone *et al.* noted issues concerning the inferential reproducibility of two studies using machine learning to increase efficiency in behavioural diagnosis of ASD [68]. They noted errors in their experimental setups and the conclusions that were drawn. In an interdisciplinary area of research such as this, Bone *et al.* suggest that increased collaboration is needed between experts in the applied field and researchers from the computational sciences [68]. If care is not taken to understand the underlying research and tools in the area of ASD nosology, invalid

conclusions may be drawn from experiments. Likewise, if an expert in the clinical sciences uses off-the-shelf ML models without properly understanding the assumptions made about the data and making the necessary adjustments, the experiments may be invalidated.

The heterogeneity of ASD is certainly a source of difficulty for this classification problem. Reiter *et al.* conducted a study comparing classifier accuracy on four subsets of the ABIDE dataset (as well as in-house data). They split the groups as follows:

1. Both sexes, unrestricted severity.
2. Male participants, unrestricted severity.
3. Both sexes, higher severity only.
4. Male participants, higher severity only.

They found that the accuracy increased when the homogeneity of the dataset increased, implying that the different severities and the sex of the individuals had a noticeable impact on how well ASD could be identified [69]. Bölte *et al.* also suggested the stratification of subgroups within ASD to develop a better understanding of its etiology [70]. As discussed earlier in Section 5.1, it might not make sense to consider this problem as a binary classification problem. Taking into account specific symptoms and their severity may help with classification and also give insights into the severity of ASD for patients being diagnosed. It should be noted that both ABIDE datasets include useful phenotypic information associated with each subject, and ABIDE II also includes various symptoms and their severities.

Moreover, though the definition of ASD was changed in the DSM-5 to include autistic disorder, Asperger's disorder, and pervasive developmental disorder not otherwise specified (PDD-NOS) [4] due to the presence of poor reliability data concerning their diagnoses [71], it might be useful to categorize patients by their DSM-IV-TR classifications which are provided with the ABIDE datasets. A potentially superior approach may be to use measures of graph characteristics and graph similarities [61] to perform unsupervised classification on the brain networks. Perhaps data-driven definitions of subcategories of ASD could be developed based on graph clustering, or perhaps the differences in severity can be detected.

Resting-state fMRI data may not be ideal for the detection of ASD. Woo *et al.* noticed the focus on resting state data in this area of research and noted possible

confounding factors [16]. For example, individuals likely exist in varying emotional states and think entirely different thoughts than others during a resting state scan. It might instead be better to develop a dataset of task-based scans, especially tasks relating to the affected areas of ASD, such as social interactions. However, this is not possible if scanning younger subjects. Additionally, researchers may consider using other measures of brain physiology such as EEG to give multiple perspectives on a subject, though this would be resource-intensive and may not be practical for gathering data or using in practice.

An idea worth exploring would be to consider the spatiality of ROIs when determining features for classification. Relión *et al.* suggest this would not be as helpful when the brain network has few ROIs, which is the case with the AAL atlas used in this study, though many alternatives exist [72].

To gain some insights into how DL models achieve better classification accuracy in this domain, it may be possible to run simplified, explainable classifiers alongside black-box classifiers, and report on the instances for which they disagree to help understand what the DL models are doing differently.

Another idea is to count the number of times each brain network gets classified correctly or incorrectly over a series of experiments. Subjects that were misclassified more often could be identified, and perhaps useful insights could come from identifying what makes them difficult to correctly classify.

Finally, a new hybrid approach based on some of the findings in this study could be tested on the ABIDE datasets. This approach could include the following steps:

- Derive features using DE on a signed effect size matrix rather than the difference network used in this study.
- Calculate additional features using measures of graph characteristics [61].
- Create a Decision Tree model using the above features for clear explainability.
- Perform the above steps for multiple atlases and then let the decision trees vote on the classification.

This model would utilize various perspectives and would be extremely transparent. The classifications made would be easily interpretable.

Chapter 6

Conclusions

This study recreated the work of Lanciano *et al.* in their paper "Explainable Classification of Brain Networks via Contrast Subgraphs", but found inconsistent results with what was claimed. Modifications to Lanciano *et al.*'s methods were made, such as using a quadratic programming solution rather than semi-definite programming to approximate the densest subgraph, improving the local search algorithm used, and finding and utilizing multiple contrast subgraphs. These modifications resulted in comparable accuracies and a significant reduction in computational runtimes.

A simpler approach was developed and named Discriminative Edges (DE). Rather than determining a subgraph defined as a set of nodes representing regions of interest (ROIs) in the brain, a set of the most discriminative connections or edges between ROIs was identified using the most positive and negative values in the difference network as constructed by Lanciano *et al.* This eliminates the unnecessary inclusion of connections that do not provide discriminative information, simplifies the calculations performed, and provides a more interpretable and explainable solution. The accuracies obtained by this approach were also comparable to the other approaches, but the computational runtime was significantly lower than all other approaches examined in this study.

Based on the results obtained in this study, it appeared that the thresholding of correlation matrices performed by Lanciano *et al.* resulted in lower classification accuracy and may have removed valuable information. Experiments were conducted to vary the threshold value, but none of the values led to improved performance over the threshold chosen by Lanciano *et al.* Though a direct comparison of the approaches using thresholded and raw correlation matrices cannot be made, it was observed that the approaches using raw correlation matrices performed better overall.

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 dataset are not adequate to be trusted in practice in most fields, let alone the medical field.

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 [4], 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.

The problem also suffers from a lack of data availability. 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, etc. 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 difficulties in obtaining data for younger subjects.

The easiest way to alleviate this issue is to perform similar preprocessing on ABIDE II as the PCP 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 needed for subjects of all ages.

This study discovered some of these challenges while attempting to replicate the work of others and advance the area of research. 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.

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 the findings and suggestions of this study will aid in future efforts to meet this goal.

Bibliography

[1] 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.

[2] 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, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, 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, pp. 659–67, 2014.

[3] G. Vandebos, *APA Dictionary of Psychology*. 01 2007.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] M. J. Maenner, K. A. Shaw, A. V. Bakian, D. A. Bilder, M. S. Durkin, A. Esler, S. M. Furnier, L. Hallas, J. Hall-Lande, A. Hudson, *et al.*, "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2018," *MMWR Surveillance Summaries*, vol. 70, no. 11, p. 1, 2021.

[10] 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.

[11] 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.

[12] 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.

[13] R. E. Frye, S. Vassall, G. Kaur, C. Lewis, M. Karim, and D. Rossignol, "Emerging biomarkers in autism spectrum disorder: a systematic review," *Annals of translational medicine*, vol. 7, no. 23, pp. 792–792, 2019.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

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

[24] 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.

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

[26] 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.

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

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

[29] N. K. Logothetis and B. A. Wandell, "Interpreting the bold signal," *Annual review of physiology*, vol. 66, no. 1, pp. 735–769, 2004.

[30] S. A. Huettel, A. W. Song, G. McCarthy, *et al.*, *Functional magnetic resonance imaging*, vol. 1. Sinauer Associates Sunderland, 2004.

[31] A. C. Evans, A. L. Janke, D. L. Collins, and S. Baillet, "Brain templates and atlases," *NeuroImage*, vol. 62, no. 2, pp. 911–922, 2012. 20 YEARS OF fMRI.

[32] K. Murphy and M. D. Fox, "Towards a consensus regarding global signal regression for resting state functional connectivity mri," *NeuroImage*, vol. 154, pp. 169–173, 2017. Cleaning up the fMRI time series: Mitigating noise with advanced acquisition and correction strategies.

[33] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. S. Khundrakpam, J. D. Lewis, Q. Li, M. Milham, *et al.*, "The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives," *Frontiers in Neuroinformatics*, vol. 7, 2013.

[34] 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.

[35] 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.

[36] M. F. Misman, A. A. Samah, F. A. Ezudin, H. A. Majid, Z. A. Shah, H. Hashim, and M. F. Harun, "Classification of adults with autism spectrum disorder using deep neural network," in *2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, pp. 29–34, IEEE, 2019.

[37] H. Abbas, F. Garberson, S. Liu-Mayo, E. Glover, and D. P. Wall, "Multi-modular ai approach to streamline autism diagnosis in young children," *Scientific reports*, vol. 10, no. 1, pp. 1–8, 2020.

[38] A. Knopf, "Fda authorizes marketing of diagnostic aid for autism spectrum disorder," *The Brown University Child & Adolescent Psychopharmacology Update*, vol. 23, no. 8, pp. 7–7, 2021.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[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, pp. 599–599, 2013.

[45] 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.

[46] 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.

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

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

[49] C. Coupette, S. Dalleiger, and J. Vreeken, "Differentially describing groups of graphs," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 3959–3967, Jun. 2022.

[50] H. Wang, Y. Deng, L. Lü, and G. Chen, "Hyperparameter-free and explainable whole graph embedding," *CoRR*, vol. abs/2108.02113, 2021.

[51] 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.

[52] 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.

[53] O. D. Balalau, F. Bonchi, T.-H. H. Chan, F. Gullo, and M. Sozio, "Finding subgraphs with maximum total density and limited overlap," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, (New York, NY, USA), p. 379–388, Association for Computing Machinery, 2015.

[54] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[55] 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.

[56] S. Guan and M. Loew, "A novel intrinsic measure of data separability," *Applied Intelligence*, pp. 1–17, 2022.

[57] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *NeuroImage (Orlando, Fla.)*, vol. 15, no. 1, pp. 273–289, 2002.

[58] E. T. Rolls, M. Joliot, and N. Tzourio-Mazoyer, "Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas," *NeuroImage (Orlando, Fla.)*, vol. 122, pp. 1–5, 2015.

[59] E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot, "Automated anatomical labelling atlas 3," *NeuroImage (Orlando, Fla.)*, vol. 206, pp. 116189–116189, 2020.

[60] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia and analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[61] A. Mheich, F. Wendling, and M. Hassan, "Brain network similarity: methods and applications," *Network Neuroscience*, vol. 4, pp. 507–527, 07 2020.

[62] E. Redcay and E. Courchesne, "Deviant functional magnetic resonance imaging patterns of brain activity to speech in 2–3-year-old children with autism spectrum disorder," *Biological psychiatry (1969)*, vol. 64, no. 7, pp. 589–598, 2008.

[63] H. E. Plesser, "Reproducibility vs. replicability: A brief history of a confused terminology," *Frontiers in Neuroinformatics*, vol. 11, 2018.

[64] S. N. Goodman, D. Fanelli, and J. P. Ioannidis, "What does research reproducibility mean?," *Science translational medicine*, vol. 8, no. 341, pp. 341ps12–341ps12, 2016.

[65] V. Bajpai, M. Kühlewind, J. Ott, J. Schönwälder, A. Sperotto, and B. Trammell, "Challenges with reproducibility," in *Proceedings of the Reproducibility Workshop*, pp. 1–4, 2017.

[66] A. L. Beam, A. K. Manrai, and M. Ghassemi, "Challenges to the reproducibility of machine learning models in health care," *JAMA : the journal of the American Medical Association*, vol. 323, no. 4, pp. 305–306, 2020.

[67] A. Di Martino, D. O'Connor, B. Chen, K. Alaerts, J. S. Anderson, M. Assaf, J. H. Balsters, L. Baxter, A. Beggiato, S. Bernaerts, L. M. E. Blanken, S. Y. Bookheimer, B. B. Braden, L. Byrge, F. X. Castellanos, M. Dapretto, R. Delorme, D. A. Fair, I. Fishman, J. Fitzgerald, L. Gallagher, R. J. J. Keehn, D. P. Kennedy, J. E. Lainhart, B. Luna, S. H. Mostofsky, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, M. Solomon, R. Toro, C. J. Vaidya, N. Wenderoth, T. White, R. C. Craddock, C. Lord, B. Leventhal, and M. P. Milham, "Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii," *Scientific data*, vol. 4, no. 1, pp. 170010–170010, 2017.

[68] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: Pitfalls and promises," *Journal of autism and developmental disorders*, vol. 45, no. 5, pp. 1121–1136, 2015.

[69] M. A. Reiter, A. Jahedi, A. R. J. Fredo, I. Fishman, B. Bailey, and R.-A. Müller, "Performance of machine learning classification models of autism using resting-state fmri is contingent on sample heterogeneity," *Neural computing & applications*, vol. 33, no. 8, pp. 3299–3310, 2021.

[70] S. Bölte, S. Girdler, and P. B. Marschik, "The contribution of environmental exposure to the etiology of autism spectrum disorder," *Cellular and molecular life sciences : CMLS*, vol. 76, no. 7, pp. 1275–1297, 2019.

[71] D. A. Regier, E. A. Kuhl, and D. J. Kupfer, "The dsm-5: Classification and criteria changes," *World psychiatry*, vol. 12, no. 2, pp. 92–98, 2013.

[72] J. D. A. Relión, D. Kessler, E. Levina, and S. F. Taylor, "Network classification with applications to brain connectomics," *The annals of applied statistics*, vol. 13, no. 3, p. 1648, 2019.

Appendix A

Reproducibility

Repositories:

- This study: `https://github.com/keanelekenns/brain-network-classification`
- "Explainable Classification of Brain Networks via Contrast Subgraphs": `https://github.com/tlancian/contrast-subgraph`