Delft University of Technology

BSc Electrical Engineering

Design and Evaluation of Classifiers for Autism Spectrum Disorder from rs-fMRI Data

– Autism Detection based on Brain Graph Features –

C.X.W. Chen & H.-Rh. Stemerdink Kakisina

June 2025

Design and Evaluation of Classifiers for Autism Spectrum Disorder from rs-fMRI Data

Autism Detection Based on Brain Graph Features

by

Carmen Xiu Wen Chen & Hannah-Rhys Stemerdink Kakisina

to obtain the degree of Bachelor of Science

Supervised by Prof. dr. ir. Geert Leus and ir. Ruben Wijnands
Proposed by Prof. dr. ir. Geert Leus
Department of Microelectronics

Defended before Prof. dr. ir. Olindo Isabella, Prof. dr. ir. Geert Leus
and ir. Ruben Wijnands

Bachelor Graduation Thesis
June 25, 2025

Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Electrical Engineering Programme

Abstract

This thesis details the implementation and evaluation of seven machine learning classifiers for the detection of Autism Spectrum Disorder (ASD) using resting-state functional MRI (rs-fMRI) data from the ABIDE I dataset. Two feature representations were compared: traditional Pearson correlation features and graph-based features. Logistic Regression (LR), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) achieved the highest performance on Pearson correlation features, satisfying all predefined non-functional requirements, with average balanced accuracies up to 64.4% (SVM) and standard deviations below 2.5%. Linear Discriminant Analysis (LDA) narrowly missed the standard deviation constraint with 0.5%.

In contrast to the Pearson correlation features, graph-based features yielded consistently lower balanced accuracies, typically ranging from 54% to 59% across classifiers, underscoring their limited informativeness in the current implementation. Feature importance analysis on Pearson correlation data revealed connections between brain regions involving the inferior occipital gyrus, middle temporal pole, precuneus, and cerebellum as consistently influential for classification.

To facilitate neuroscientific exploration, an interactive tool, NASDA (Neuroimaging Autism Spectrum Disorder Analyser), was developed and demonstrated to fulfil all functional and non-functional requirements for Pearson correlation based analysis using LR as the recommended classification model.

These results highlight the dependency of classifiers performance on the quality of input features and contribute to ongoing efforts to localise robust neurological biomarkers for ASD.

Preface

This thesis was written as part of the Bachelor Graduation Project. Our group was tasked with developing a machine learning pipeline to support the identification of Autism Spectrum Disorder (ASD) based on brain imaging data. The goal was not only to build a functioning classifier, but also to gain insight into which brain connectivity features contribute most to distinguishing individuals with ASD.

Together with four other group members, we developed a tool that performs ASD classification and highlights relevant features. The overall pipeline consists of three key components: feature extraction, classification, and feature selection, each handled by a dedicated subgroup.

Our subgroup focused on the design, implementation, and evaluation of a range of classifiers, as well as the integration of all subsystems. Additionally, we conducted feature importance analysis to help bridge model output with neuroscientific interpretation.

We would like to express our gratitude to our supervisors Prof. Dr. ir. Geert Leus and ir. Ruben Wijnands for their guidance during this project. We would also like to thank ir. Dimme de Groot for his valuable feedback.

Carmen Chen & Hannah-Rhys Kakisina
June 2025

Contents

1 Introduction

According to Statistic Netherlands (CBS), approximately 3% of the Dutch population reported having Autism Spectrum Disorder (ASD) in the 2022-2024 period [1]. ASD is a complex condition that affects brain development. People with ASD experience challenges with social interaction, communication, and may engage in repetitive behaviours, often referred to as "stimming". The presentation of these traits vary widely, making ASD a heterogeneous condition. Comorbid mental health issues are also highly prevalent in the ASD population. Among adults with ASD, the pooled estimates indicated a 27% current and 42% lifetime prevalence for anxiety disorder, and 23% current and 37% lifetime prevalence for depressive disorder [2]. These elevated rates highlight the importance of early diagnosis to ensure that individuals with ASD can receive proper support to improve their quality of life.

Currently, ASD diagnosis primarily relies on behavioural assessments based on criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [3]. Although this is effective in many cases, it may fall short when symptoms overlap with other conditions. As a result, there is a growing interest in identifying neurobiological markers of ASD using non-invasive brain imaging techniques.

A promising technique is resting state functional Magnetic Resonance Imaging (rs-fMRI), which measures brain activity while the subject is not engaged in any specific task. From this data, one can compute functional connectivity (FC), which is defined as the correlation between the blood-oxygen-level-dependent (BOLD) signal time series from different brain regions. FC quantifies how strongly different areas of the brain are functionally linked. It has been hypothesised that individuals with ASD exhibit altered patterns of FC [4], which could potentially serve as a basis for identifying neural biomarkers capable of distinguishing between individual with ASD and non-autistic (allistic) individuals.

To explore these potential biomarkers, our study uses machine learning methods to classify between individuals with ASD and allistic/typical control (TC) individuals. The models are then analysed to identify which brain regions contributed most strongly to their predictions. These regions may offer insight into the underlying neurobiology of ASD and highlight potential biomarkers for further neuroscientific research.

A key challenge in neuroimaging-based ASD research has been the limited availability of a large-scale rs-fMRI dataset with reliable diagnostic labels. To address this, the Autism Brain Imaging Data Exchange I (ABIDE I) initiative was launched in 2014 [5]. ABIDE I aggregated existing rs-fMRI data from 17 international research sites to facilitate a large-scale rs-fMRI dataset.

The various acquisition sites differ in scanner hardware, acquisition protocols, and participant demographics. Although this initiative dramatically increased the availability of labelled rs-fMRI data for ASD research, it also introduced a significant degree of heterogeneity. This heterogeneity poses a major challenge for machine learning models, as it becomes difficult to determine whether the observed patterns in the data reflect true neurodevelopmental differences related to ASD or site-specific confounding effects.

1.1 Research Goals and Scope

The aim of our part of the research is to build reliable models responsible for the classification task. From the classification we can trace back which brain regions or connections between them were important in the decision-making. In doing so, a potential neural biomarker can be found. The selection of these models will be discussed in Section 3.4. The performance of the models applied will be evaluated based on accuracy, interpretability and generalisability and are aimed to satisfy the non-functional requirements discussed in Chapter 2.

In addition, we investigate two other methods to assist rs-fMRI based ASD research. Firstly, the contribution of graph-based features to the model performance will be assessed in comparison to the traditional features. Secondly, we will deliver an interactive tool to control the pipeline and visualise the regions of interest (ROIs) in an intuitively and interpretable manner.

This research investigates both single-site and multi-site classification of ASD as both come with their own challenges. For single-site classification, the limited data availability poses a real challenge, whereas multi-site classification is more difficult due to the site heterogeneity.

1.2 Related Work

Machine learning applied to ASD classification using rs-fMRI has progressed steadily over the past decade. Table 1.1 summarises key studies, focussing on dataset size, feature types, feature selection methods, classifiers, and performance.

Table 1.1: A summary of studies on Autism Spectrum Disorder (ASD) multi-site classification using resting-state functional MRIs (rs-fMRIs) and machine learning. Each study is characterised by the number of ASD and typical control (TC) participants, the type of features used (functional connectivity or graph-based), feature selection methods (if any), machine learning classifiers employed, and the reported classification accuracy. Functional connectivity features were typically computed between brain regions defined by parcellation atlases, most commonly the Automated Anatomical Labelling (AAL) and Craddock 200 (CC200) atlases [16], [17]. The studies span a range of approaches from traditional interpretable models to deep learning. Deep learning approaches generally achieved higher classification accuracy, but often involve greater computational cost and reduced interpretability. All studies used the ABIDE I dataset.

	Participants	Features	Feature Selection	Classifier	Accuracy
[6]	447 ASD / 517 TC	FC between 7266 ROIs	None	Not specified	60%
[7]	403 ASD / 468 TC	FC between ROIs (3 atlases)	ICA, MSDL	SVC, ridge regression	66.8%
[8]	505 ASD / 530 TC	FC between 200 ROIs (CC200)	DAE	DNN classifier	70%
[9]	505 ASD / 530 TC	FC between ROIs (AAL)	None	Ridge, LR, l-SVM, k-SVM	65.4–66.2%
[10]	505 ASD / 530 TC	FC between 200 ROIs (CC200)	AE	Single-layer perceptron	70.3%
[11]	403 ASD / 468 TC	Laplacian eigenvalues and centrality	SBS	LR, LDA, KNN, SVM, NN	71.7-77.7%
[12]	506 ASD / 548 TC	FC between 200 ROIs (CC200)	Extra-trees	l-SVM	72.2%
[13]	505 ASD / 530 TC	FC between 116 ROIs (AAL)	None	DNN classifier	74%
[14]	493 ASD / 530 TC	Graph features	PCA	Multilayer perceptron	64.4%
[15]	306 ASD / 350 TC	FC between 237 ROIs	CRF	RF	62.5%

Abbreviations: TC: typical control; ROI: region of interest; AAL: Automated Anatomical Labeling; FC: functional connectivity; SVM: Support Vector Machine; DAE: denoising auto encoder; DNN: deep neural network; AE: auto encoder; SBS: sequential backward feature selection; PCA: principal component analysis; ICA: independent component analysis; MSDL: multi-subject dictionary learning; LR: logistic regression; CRF, conditional random forest.

Early work by Nielsen *et al.* has shown the feasibility of using FC for ASD classification using the ABIDE I dataset, achieving an accuracy of approximately 60% through leave-one-out cross validation. However, the specific classifiers used in their study were not detailed [6].

Without using feature selection, Yang *et al.* assessed the performance of several interpretable models, including ridge classifier, Logistic Regression (LR), linear Support Vector Machine (l-SVM), and kernel SVM (k-SVM) on FC features derived from the AAL atlas. They obtained accuracies ranging from 65.4% to 66.2%, with their LR model yielding the best accuracy [9].

To overcome the performance limitations of interpretable models, several studies have adopted deep learning. Heinsfeld *et al.* implemented denoising auto encoders (DAEs) to learn latent representations of the full connectivity features extracted from 200 ROIs (19900 features). These were used as input to an MLP classifier. Their method produced an accuracy of 70% on the ABIDE I dataset. However, the reported training time for their model was approximately 33 hours [8]. Similarly, Eslami *et al.* proposed their model called ASD-DiagNet, where they applied autoencoding and used a single layer perceptron. They used a data augmentation technique inspired by Synthetic Minority Over-sampling Technique (SMOTE) to create enough data for their model, to prevent it from overfitting. Their proposed model achieved an accuracy of 69.4% without augmentation and an accuracy of 70.3% with augmentation on the CC200 atlas. They reported a running time of 41 minutes for ASD-DiagNet with augmentation, and a running time of 20 minutes without augmentation [10].

Recent work has explored graph-theoretical features. Mostafa *et al.* extracted Laplacian eigenvalues from 264 regions and three features by network centralities. With backward sequential feature selection, 64 features were selected. Their highest obtained accuracy was 77.7% with the LDA model [11]. Kazeminejad and Sotero discussed the role of anti-correlation in graph-theory based ASD classification and used graph features with a multilayer perceptron achieving 64.4% accuracy [14].

Liu *et al.* used feature selection via Extra-Trees to train an l-SVM on FC features derived from the CC200 atlas, achieving 72.2% accuracy [12]. Reiter *et al.* examined the effect of heterogeneity on classification performance. FC features were extracted from 237 ROIs and conditional random forests was implemented for feature selection in combination with an RF classifier. Their approach achieved an accuracy of 62.5% on full

heterogeneity and 65% on reduced gender heterogeneity [15].
Abraham *et al.* evaluated ICA and MSDL and achieved 66.8% accuracy using SVMs and ridge regression [7]. These studies collectively show that interpretable models without feature selection or harmonisation typically yield accuracies in the 60–66% range. Accuracy can be improved using deep learning or feature selection, though often at the cost of interpretability and running time. Graph features leveraging topological metrics show promising performance and neuroscientific relevance. Lastly, multisite heterogeneity continues to limit generalisability, requiring harmonisation or domain adaptation strategies.

1.3 Research Questions

To structure our study, we address the following research questions:

1. Which classifiers and feature types offer the best trade-off between accuracy and interpretability?
2. How can we identify brain regions that are both consistently important in the prediction across classifiers and experimental folds and how are these identified regions supported by established neuroscientific findings in ASD?
3. How can we develop an intuitive interface to control the pipeline and visualise the resulting ROIs and model performance to support further neuroscientific analysis?
4. How can we improve the generalisability of our classifiers?
5. What evaluation methods best capture meaningful classifier performance beyond average accuracy, including site-specific, sex-based, or age-based assessments?

1.4 Thesis Outline

This thesis is organised as follows. Chapter 2 presents the programme of requirements. Chapter 3 details the methodology, including data preprocessing, feature extraction, classification methods, evaluation strategies, model interpretation, and lastly the graphical user interface. Chapter 4 focusses on the results from the single-site classification experiments on Pearson correlation features. Chapter 5 addresses the multi-site classification on Pearson correlation features, discussing harmonisation and generalisation across sites. Chapter 6 focusses on the feature importance analysis on Pearson correlation features. Chap 7 focusses on the results from multi-site classification on graph-based features designed by the Feature Design subgroup. Chapter 8 introduces our graphical user interface. In Chapter 9 we answer the research outlined in Section 1.3. Lastly, Chapter 10 concludes this research with final remarks and directions for future research.

2 Programme of Requirements

Our tool will be used for neuroscientific research to find potential biomarkers. To offer reliable results, we have defined functional and non-functional requirements for our system. The performance metrics we want to obtain with our classifiers are based on what has been achieved in previous studies [6], [9], [14].

Functional Requirements

FR1 The system must classify between allism and autism

FR2 The system must highlight the brain regions involved in the decision-making.

FR3 The system must use the ABIDE dataset.

FR4 The system must allow users to select different features.

FR5 The system must allow users to select different classifiers.

FR6 The system must allow users to select different graph inference algorithms.

FR7 The system must provide performance metrics.

Non-Functional Requirements

NFR1 The system must achieve a minimum accuracy of 62% before feature selection.

NFR2 The system must achieve a minimum sensitivity of 58% before feature selection.

NFR3 The system must achieve a minimum specificity of 65% before feature selection.

NFR4 The system must achieve a minimum AUROC of 0.65 before feature selection.

NFR5 The system must achieve a minimum balanced accuracy of 62% before feature selection.

NFR6 The standard deviation (SD) of each performance metric across 5-fold stratified cross-validation must not exceed 5 percentage points.

NFR7 The computation time of a single execution of 5-fold cross-validation should stay below 40 minutes.

NFR8 The system must be open source.

NFR9 The system must be intuitive.

3 Methodology

In this research the Autism Brain Imaging Data Exchange I (ABIDE I) dataset is utilised for testing and training [5]. The dataset contains 1112 subjects in total. Tables A.1 and A.2 in Appendix A contain subject information of the ABIDE I dataset and the scanners used by the acquisition sites, respectively.

3.1 Neuroimaging Data Preprocessing

The ABIDE I dataset was preprocessed using the Configurable Pipeline for the Analysis of Connectomes (CPAC) [18], a standardised pipeline that includes motion correction, skull stripping, intensity normalisation, nuisance regression, band-pass filtering, and spatial normalisation. See Appendix A.1 for the complete list of preprocessing steps.

After preprocessing and filtering (due to motion threshold, missing information, phenotypic file issues, and quality assurance), the dataset was reduced from 1112 to 871 usable subjects. This reduction is consistent with previous studies using ABIDE I data, which similarly report substantial loss of subjects due to preprocessing [7], [11]. See Appendix A.2 for exact counts and exclusion rationale. Table A.3 in Appendix A provides the demographic characteristics of the resulting cohort after preprocessing.

Our preprocessing choices were influenced by the need for consistency with the Feature Design subgroup, whose features were derived using bandpass filtering (BPF) without global signal regression (GSR) [19]. To ensure a fair comparison between Pearson correlation and graph-based features, we adopted the same preprocessing pipeline. This decision was made to prevent the confounding effects that different preprocessing steps might introduce. In this manner, we can maintain the integrity of our comparative analyses.

3.2 Feature Extraction

After preprocessing, two types of features were extracted for the classification of ASD. First, we extracted Pearson correlation features, which serve as a baseline in this study. Second, the feature design subgroup developed graph-based features. These two were compared to assess whether graph features provide added discriminative power beyond conventional correlation based features.

3.2.1 Pearson Correlation Features

In Chapter 4 to Chapter 6, we will focus on the classification of ASD using Pearson correlation features. This is a way to represent functional connectivity (FC) in rs-fMRI studies. These features are widely used due to their simplicity and proven effectiveness in previous ASD studies [8], [9]. We computed the Pearson correlation between the mean blood-oxygen-level-dependent (BOLD) time series of every pair of regions of interest (ROIs). The brain was parcellated using the Automated Anatomical Labeling (AAL) atlas, which defines 116 ROIs [20, 16]. Pearson correlation measures the linear relationship between the time series of two brain regions. In FC research, it is commonly assumed that if two regions are temporally coactivated, they may be functionally connected [21]. This results in a symmetric FC matrix, from which we extracted $\binom{116}{2} = 6670$ unique features per subject by taking the upper triangle (excluding the diagonal). To preserve interpretability, the features were named based on the AAL region indices of the connected ROI pair. These indices were obtained using `nilearn`'s `fetch_aal_atlas()` [22] and can be mapped back to anatomical labels. For example, the index `2001` corresponds to the region "Precentral_L". This mapping facilitates the identification of important brain connections during feature importance analysis, thereby supporting the system's ability to highlight the brain regions involved in the classification as required by **FR2**. A complete list of AAL index-to-label mappings is provided in Appendix B.

3.2.2 Graph Features from Feature Design

In Chapter 7, we will focus on the classification of ASD using graph-based features provided by the Feature Design subgroup. These features were extracted by first applying Smith Independent Component Analysis (ICA) or group ICA to reduce the dimensionality of the rs-fMRI data. Then, based on those ICA components, a variety of graph inference techniques were used to construct brain connectivity graphs [19]. These included both statistical inference methods and graph signal processing (GSP)-based approaches.

The statistical methods consisted of sample covariance, Pearson correlation, partial correlation, mutual information, and Granger causality. The GSP-based approaches included `rSpecT` (normalized Laplacian and adjacency) and `rLogSpecT`, a variation that includes a log barrier. The exact distinction between these methods, along with the overall feature extraction process, is detailed in the thesis of the Feature Design subgroup [19]. Most graph inference methods relied on the estimation of the sample covariance matrix as input, with the exception of Pearson correlation, mutual information, and Granger causality. To this end, seven covariance estimation methods were implemented: direct empirical estimation, sliding-window averaging, Ledoit-Wolf shrinkage, graphical Lasso (Glasso), time-varying Glasso, vector autoregression (VAR(k)), and nonlinear VAR (NVAR).
In total, we received 29 different feature matrices generated from various combinations of inference methods and covariance estimation techniques. To systematically manage these datasets and clearly indicate the inference configuration used to generate them, we assigned a unique identifier to each feature matrix. Each matrix has a shape of $(n \times m)$, where n denotes the number of samples (subjects) and m the number of features, where all features are scalars. In addition, the specific features varied per combination of ICA, graph inference, and covariance estimation methods applied. A detailed overview of the ICA, graph inference, covariance estimation methods, and corresponding features for each matrix is provided in Table C.1 in Appendix C.2.

3.3 Feature Preprocessing

Once the features were extracted, we performed the preprocessing steps necessary for classification. This includes mean imputation for the missing values and feature standardisation by removing the mean and scaling to unit variance.

3.3.1 Harmonisation

One of the key challenges in multi-site classification is the heterogeneity introduced by variations in acquisition protocols, population demographics, and scanner hardware. This can lead to site-specific biases in the extracted features, which in turn can lower model performance and hinder generalisability across acquisition sites [23]. To mitigate these undesirable site effects, we explored statistical harmonisation using the `NeuroHarmonize` Python package [24], which is based on ComBat, a batch effect correction tool that removes inter-site technical variability, while preserving inter-site biological variability [25]. ComBat has also been successfully applied to neuroimaging [26]. In our study, harmonisation was applied to the Pearson correlation features prior to classification, in particular for the multi-site classification settings. `NeuroHarmonize` adjusts for site differences by fitting a location-scale model that estimates and removes additive and multiplicative site effects [24], [27]. Previous work has demonstrated that ComBat-based harmonisation can enhance classifier performance in related applications, such as rs-fMRI-based detection of Major Depressive Disorder (MDD) [28]. However, a limitation of ComBat-based harmonisation is that it requires all sites present in the test set to also be observed during training. To satisfy this assumption, we excluded the CMU site from our experiments due to its low number of subjects, which made it likely to be omitted from training folds during cross-validation.

Harmonisation was applied exclusively to the Pearson correlation features. The graph-based features provided by the Feature Design subgroup were not harmonised, as they were derived from a complex pipeline involving dimensionality reduction, diverse graph inference methods, and graph-theoretic transformations. As a result, the extracted features represent higher-level abstractions, rather than direct measurements from the raw rs-fMRI signals. Since ComBat-based harmonisation assumes additive and multiplicative site effects in the original input features, its application to these graph features would not be appropriate.

3.4 Classification methods

To detect ASD patterns from functional brain features, reliable classification models are needed. We will discuss all classification methods used in this research below. Given the complexity of the underlying brain connectivity patterns, a range of both linear and non-linear classifiers was considered. Each model was implemented using the `scikit-learn` library [29], and where appropriate, tuned to optimise performance. The following subsections detail the mathematical foundations, practical implementation, and hyperparameter configurations of each classifier: Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees

(DT), Random Forests (RF), Multi-Layer Perceptrons (MLP), Linear Discriminant Analysis (LDA), and K-Nearest Neighbours (KNN).
For all classifiers, we let the training samples be represented as $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, where $x_i \in \mathbb{R}^n$ and $y_i \in \{0, 1\}$, the vector $w \in \mathbb{R}^n$ represent the weights and $w_0 \in \mathbb{R}$ the bias of the model.

3.4.1 Logistic Regression

Logistic Regression is a linear classifier widely used in binary classification tasks, including ASD classification [9], [11]. In this research, regularised Logistic Regression has been applied with an L2 penalty to prevent overfitting. The model solves the following optimisation problem [29]:

$$\begin{aligned} \min_{w} \quad & \frac{1}{S} \sum_{i=1}^{m} s_i \left[-y_i \log\left(\hat{p}\left(x_i\right)\right) - \left(1 - y_i\right) \log\left(1 - \hat{p}\left(x_i\right)\right) \right] + \frac{1}{2SC} \|w\|_2^2 \\ \text{s.t.} \quad & \hat{p}\left(x_i\right) = \frac{1}{1 + \exp\left(-x_i^T w - w_0\right)} \\ & S = \sum_{i=1}^{m} s_i, \end{aligned} \tag{3.1}$$

where C is the inverse regularisation strength and s_i corresponds to the weights assigned to a specific training sample (the vector $\mathbf{s}$ is formed by element-wise multiplication of the class weights and sample weights). In our model, we did not apply any class or sample weighting. All classes were assigned equal importance (class weight set to one) and all training samples were treated equally (unit sample weights).

This optimisation problem was solved using the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [30], which is well suited for high-dimensional problems due to its low memory usage and fast convergence properties.

3.4.2 Support Vector Machine

Support Vector Machines are supervised learning models that aim to find the optimal separating hyperplane between two classes in a high-dimensional feature space [31]. The optimal hyperplane is the one that maximises the margin between the classes, defined as the distance to the nearest support vectors of said classes. The SVM solves the following optimisation problem [29]:

$$\begin{aligned} \min_{w,\zeta} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{m} \zeta_i, && i = 1, \ldots, m \\ \text{s.t.} \quad & y_i (w^T \phi(x_i) + w_0) \geq 1 - \zeta_i, && i = 1, \ldots, m \\ & \zeta_i \geq 0, && i = 1, \ldots, m \end{aligned} \tag{3.2}$$

where C acts as an inverse regularisation parameter that controls the trade-off between maximising the margin and penalising classification errors, $\phi(x_i)$ denotes the implicit mapping to a higher-dimensional space induced by the kernel functions, given in (3.3)–(3.6) and ζ_i is the maximum distance of samples from their correct margin boundary allowed if not all samples can be correctly separated [29].

Hyperparameter Tuning for SVM

For the SVM given by (3.2), we evaluated kernel types linear (3.3), radial basis function (RBF) (3.4), polynomial (3.5), and sigmoid (3.6) [29] using hyperparameter tuning via a randomised grid search. Key hyperparameters included the regularisation parameter C, the kernel-specific parameter γ, the degree d, and the kernel coefficient r.

$$\text{Linear:} \quad \phi(x) = \langle x, x' \rangle \tag{3.3}$$

$$\text{RBF:} \quad \phi(x) = \left(\gamma \langle x, x' \rangle + r\right)^d \tag{3.4}$$

$$\text{Poly:} \quad \phi(x) = \exp\left(\gamma \langle x, x' \rangle + r\right)^d \tag{3.5}$$

$$\text{Sigmoid:} \quad \phi(x) = \tanh\left(\gamma \langle x, x' \rangle + r\right) \tag{3.6}$$

where $x' \in \mathbb{R}^n$ is your test sample that gets compared to all x from the training set by the kernel function using the "kernel trick" [32].

3.4.3 Tree Based Methods

Both Decision Tree and Random Forest were incorporated in our analysis. Decision Trees are non-parametric models that recursively partition the feature space to learn decision boundaries.

Let $Q_l \in \mathbb{R}^{n_l}$ represent the subset of training vectors at a node l, with n_l samples. If for each candidate split $\theta = (j, t_l)$, consisting of a feature j and threshold t_l, partitions the subset into the new $Q_l^{\text{left}}(\theta)$ and $Q_l^{\text{right}}(\theta)$ subsets, then the Decision Tree can be optimised with the optimisation problem below [29].

$$
\begin{aligned}
\min_{\theta} \quad & G(Q_l, \theta) = \frac{n_l^{\text{left}}}{n_l} H\left(Q_l^{\text{left}}(\theta)\right) + \frac{n_l^{\text{right}}}{n_l} H\left(Q_l^{\text{right}}(\theta)\right) \\
\text{s.t.} \quad & Q_l^{\text{left}}(\theta) = \{(x, y) \mid x_j \leq t_l\} \\
& Q_l^{\text{right}}(\theta) = Q_l \setminus Q_l^{\text{left}}(\theta) \\
& \theta = (j, t_l),
\end{aligned}
\tag{3.7}
$$

where $H(\cdot)$ is the applied Gini or Entropy function and n_l^{left} and n_l^{right} refer to the number samples of the resulting $Q_l^{\text{left}}(\theta)$ and $Q_l^{\text{right}}(\theta)$ subsets, respectively. Decision Trees are highly interpretable, yet prone to overfitting if not tuned accordingly, especially for high-dimensional problems.

Hyperparameter Tuning for Decision Tree

To prevent overfitting we tuned our tree's hyperparameters `max_depth`, `min_samples_split`, `min_samples_leaf` and the `criterion`. The `max_depth` regulated the maximum depth of the tree, the `min_samples_split` and `min_samples_leaf` regulated the minimum number of samples required to cause a split or leaf. The `criterion` refers to the $H(\cdot)$ applied and is tuned to either the Gini (3.8) or the entropy (3.9) [29].

$$H(Q_l) = \sum_k p_{lk}(1 - p_{lk}), \tag{3.8}$$

$$H(Q_l) = -\sum_k p_{lk} \log(p_{lk}), \tag{3.9}$$

where

$$p_{lk} = \frac{1}{n_l} \sum_{y \in Q_l} I(y = k). \tag{3.10}$$

Random Forest

With the Random Forest model we aimed to reduce variance with a possible slight bias trade-off by introducing randomness. The Random Forest model is an ensemble of a large number of trees that averages the predictions to reduce overfitting. All trees were modelled to randomised sets of samples and features, to create a large variety of trees.

In accordance with the performance constraint defined in requirement **NFR7**, which states that the computation time for a single execution of 5-fold cross-validation must not exceed 40 minutes, hyperparameter tuning for the Random Forest model was omitted. Instead, the default parameter settings provided by `scikit-learn` were employed to ensure conformance to the specified time constraint. These defaults include: `n_estimators=100`, `criterion='gini'`, `max_depth=None`, `min_samples_split=2`, `min_samples_leaf=1`, `max_features='sqrt'`.

3.4.4 Multi-Layer Perceptron

A Multi-Layer Perceptron (MLP) is a neural network capable of modelling non-linear decision boundaries. It consists of one or more hidden layers of neurons, where each neuron performs a transformation followed by a non-linear activation function. For our MLP in most cases, a single hidden layer was used, making it the most simplistic form of MLP while still capable of learning non-linear patterns in the data. However, in a few rare cases, hyperparameter tuning did result in a double hidden layer. The one-layer MLP learns the function [29]

$$f(x) = \hat{w} g(w^T x + w_0) + \hat{w}_0, \tag{3.11}$$

where $w \in \mathbb{R}^n$ and $\hat{w} \in \mathbb{R}$ represent the weights of the input and hidden layers, respectively, and $w_0, \hat{w}_0 \in \mathbb{R}$ are the biases added to the hidden layer and output layer, respectively. The function $g(\cdot) : \mathbb{R} \to \mathbb{R}$ represents the activation function, which is set either to the hyperbolic tangent (3.12) or the Rectified Linear Unit (ReLU) function (3.13) [29].

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{3.12}$$

$$g(z) = \max(0, z) \tag{3.13}$$

Hyperparameter Tuning for MLP

We tuned `scikit-learn`'s `MLPClassifier`'s number of neurons, activation function, regularisation term `alpha` and initial learning rate `learning_rate_init`. The `solver` was also varied between the options `sgd` and `adam`. For the hidden layers, the tuner compared between a layer of 50 or 100 neurons or the use of both layers sequentially.

3.4.5 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a classification method derived from a probabilistic generative model. According to [29], it models the conditional density $P(x \mid y = k)$ for each class k as a multivariate Gaussian distribution. Using Bayes' theorem, the posterior probability of a class given a sample $x \in \mathbb{R}^n$ is given by:

$$P(y = k \mid x) = \frac{P(x \mid y = k)\, P(y = k)}{P(x)} \tag{3.14}$$

The prediction $\hat{y}$ is made by assigning x to the class k that maximises this posterior probability. Thus:

$$\hat{y} = \arg\max_k \, P(y = k \mid x) = \arg\max_k \, P(x \mid y = k)\, P(y = k). \tag{3.15}$$

For LDA, the class-conditional densities are modelled as multivariate Gaussians with the assumption that all classes share the same covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma}$:

$$P(x \mid y = k) = \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_k)^T \mathbf{\Sigma}^{-1} (x - \mu_k) \right) \tag{3.16}$$

which gives

$$\log P(y = k \mid x) = -\frac{1}{2} (x - \mu_k)^T \mathbf{\Sigma}^{-1} (x - \mu_k) + \log P(y = k) + \text{Cst}, \tag{3.17}$$

where the constant term Cst corresponds to the constants from the Gaussian that are neglectable for the maximisation of the log posterior. The log posterior (3.17) is equivalent to a linear function of x, and can be rewritten [33] as:

$$\log P(y = k \mid x) = w_k^T x + w_{k0} + \text{Cst}, \tag{3.18}$$

where

$$w_k = \mathbf{\Sigma}^{-1} \mu_k, \quad w_{k0} = -\frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log P(y = k).$$

Because of the shared covariance assumption, LDA yields linear decision surfaces and is therefore highly interpretable. For our LDA we implemented `scikit-learn`'s `LinearDiscriminantAnalysis` with default settings. Because the model has only few hyperparameters, no tuning was required.

3.4.6 K-Nearest Neighbours

The k-Nearest Neighbours algorithm is an instance-based learning method using majority voting, therefore there is no general internal model. A sample is assigned the most common label among its k nearest neighbours from the training set, where k is a predefined integer, set to 5. Larger values of k could be used to suppress noise but compromise strict boundaries [29].

The simplicity and interpretability of KNN make it an appealing classifier, especially when the decision boundary is expected to be non-linear. However, its performance can degrade in high-dimensional spaces due to the curse of dimensionality, and its classification time scales with the size of the training set.

3.5 Evaluation strategies

3.5.1 Stratified 5-fold cross validation

To evaluate the performance of the classifiers in a statistically robust and reproducible way, we used stratified 5-fold cross-validation (CV) throughout most of our experiments. In this strategy, the dataset is divided into five folds, such that each fold maintains the same class distribution (i.e., proportion of ASD and TC) as the full dataset. Each fold is used once as a test set, while the remaining four serve as training data. We report the average accuracy across folds using multiple metrics: accuracy, sensitivity, specificity, balanced accuracy, and the area under the receiver operating characteristic curve (AUROC). See Appendix D for formal definitions of these metrics.

To establish a baseline for classifier performance, we implemented a dummy classifier using `scikit-learn`'s `DummyClassifier` with the `strategy='stratified'` option. This classifier generates predictions that respect the class distribution of the training set but does not make use of the input features. Comparing our models against this baseline allows us to determine whether they have learned meaningful patterns or whether they are merely exploiting imbalance in the dataset. Performance metrics were computed using the same cross-validation setup and reported for overall performance only. Subgroup-specific evaluations (described below) were not applied to the dummy classifier, as its predictions are not based on input data and would not yield meaningful subgroup-specific insights.

To further understand the behaviour of the models across relevant subpopulations, we included subgroup-specific evaluations:

- **Per-site**: evaluates how performance varies across the different acquisition locations. This may help in identifying potential site-overfitting or domain shifts.
- **Per-sex**: accounts for the known sex-based differences in ASD presentation. This may help in detecting classifier bias and evaluating fairness.
- **Per-age**: since neurodevelopment changes with age, we grouped participants into meaningful age bins. Namely, 0–11, 12–18, 19–30, and 30+. We evaluated the performance within each group. These groups reflect key stages in brain development, especially within the context of ASD. This choice was supported by Guo *et al.*, who showed that individuals with ASD exhibit atypical developmental trajectories of local spontaneous brain activity across childhood (0–11), adolescence (12–18), and adulthood (19+) [34]. We further divided the adult group into 19–30 and 30+ based on the findings by Cao *et al.*, who demonstrated that global topological properties of the functional connectome, specifically the local efficiency, continue to evolve through young adulthood and shows signs of reorganisation in the early 30s [35].

As some sites and age groups contain a low number of samples, there existed a possibility that one of the two classes (ASD or TC) was absent in a certain fold during subgroup-specific evaluation of those subgroups. If one of the classes is absent, the performance cannot be calculated. For such cases, we recorded a NaN value and during averaging across folds only averaged over the folds where performance metrics were calculated.

3.5.2 Leave-One-Group-Out (LOGO) cross validation

In addition to stratified CV, we used LOGO cross-validation to test the generalisation capability of the models to unseen imaging sites. In this method, each group corresponds to an acquisition site. In each iteration, one site is held out for testing and the other sites are used during training. This simulates a real-world scenario, where a model trained on data from certain institutions must generalise to data from unseen sites, which

potentially has different scanners, protocols, and/or population characteristics. This method provides a rigorous test of generalisation, which stratified CV cannot capture.
NeuroHarmonize was not applied in LOGO as they require the test site to be present in the training set, which by definition is not the case in this setup.

3.6 Model interpretation

Feature importance analysis was conducted to identify the brain regions most influential in the model's decision-making. This analysis was restricted to the multi-site classification setting, as it demonstrated greater stability across cross-validation folds compared to single-site configurations. Specifically, the multi-site setting was selected, because it better satisfied the stability requirement in **NFR6**, which mandates that the standard deviation of performance metrics must not exceed 5 percentage points. In contrast, the single-site models exhibited greater variability and were therefore excluded from interpretation.
Interpretation was further limited to the Logistic Regression and SVM classifiers, as these were the only models that met the criteria **NFR1**, **NFR2**, **NFR3**, **NFR4**, **NFR5**, and the aforementioned **NFR6**.
To identify the most influential features, we analysed the model weights learned during training by the Logistic Regression and SVM classifiers. These weights, also called coefficients, reflect how strongly each feature contributes to the model's decision. A positive weight increases the likelihood of the model predicting ASD, while a negative weight favours the control class.
We ranked the features by the absolute value of their weights to determine importance, and retained the sign to indicate the direction of influence. The top 20 ranked features were then visualised as brain region pairs using `nilearn` for each fold in both classifiers. In addition, we visualised features that appeared in the top 20 ranking in more than 6 of the 10 folds (5 for Logistic Regression, 5 for SVM), averaging their weights over the folds in which they occurred.

3.7 Graphical User Interface

Lastly, we built NASDA (Neuroimaging ASD Analyser), a graphical user interface (GUI), that serves as a tool for neuroscientists to guide them in their research on neural biomarkers. We implemented the classifiers and connected our subsystem to the subsystems built by the Feature Design and Feature Selection subgroups. The main structure of the GUI has been built using the `tkinter` [36] and Pillow [37] Python libraries. To keep the interface modular, the full code has been subdivided in clear functions that allow multiple optional inputs. A settings option to adjust the users path to their graph features file has been incorporated in the GUI and an option to import your own data from a CSV file can be activated through the main source file. More on the exact functioning of the GUI will be discussed in Chapter 8 and its documentation can be found in Appendix G.

4 Single-Site Performance on Pearson Correlation

This chapter presents and evaluates the performance of various classifiers applied to data from ABIDE I largest site, New York University Langone Medical Center (NYU), in recognising ASD using only Pearson correlation rs-fMRI features. This site contained 74 ASD and 98 TC samples. We have singled out NYU, as its size made it least likely to make the model overfit due to data constraints, and limiting the analysis to a single acquisition site avoids inconsistencies caused by differences in scanning protocols, preprocessing, or site-specific demographics. This dataset contained 74 ASD samples of which 13.6% female and 98 control samples of which 26.5% female.

4.1 Performance evaluation

To assess model performance, 5-fold cross-validation was conducted independently three times. In each case, the models were both trained and evaluated on the same demographic subset: once on the full dataset including both female and male samples (referred to as the *combined* dataset), once on only female samples, and once on only male samples. For each configuration, we report the average sensitivity, specificity, area under the ROC curve (AUROC), accuracy, and balanced accuracy. These results are shown in Tables 4.1, 4.2, and 4.3, respectively.

4.1.1 Performance on Combined Data

The results in Table 4.1 show notable differences in performance for single-site classification on combined data across our seven classifiers. LR and MLP performed the strongest overall, with LR achieving the highest AUROC (74.1% ±6.2%) and accuracy (71.5% ±5.2%), while MLP yielded the highest sensitivity of 66.1% ±5.8% compared to the LR's sensitivity of 58.0% ±11.4%. This suggests that LR maintained a more balanced trade-off between true positive and true negative rates, whereas MLP prioritised correctly identifying ASD cases. In comparison to the dummy classifier, DT performed worse across all evaluated metrics. In terms of sensitivity, only LR and MLP performed better than the dummy classifier. For the remaining performance metrics, all models, except DT, outperformed the dummy classifier.

Table 4.1: The average performance across folds on the combined data from the NYU site using Pearson correlation features.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
Dummy	56.9% ± 6.9%	56.3% ± 13.0%	54.5% ± 5.3%	56.5% ± 7.8%	56.6% ± 7.2%.
LR	58.0% ± 11.5%	81.8% ± 8.9%	**74.1% ± 6.2%**	**71.5% ± 5.2%**	**69.9% ± 5.7%**
SVM	48.4% ± 13.8%	68.5% ± 22.7%	62.2% ± 16.1%	59.8% ± 16.9%	58.5% ± 16.3%
DT	41.7% ± 7.9%	52.0% ± 17.5%	48.6% ± 10.9%	47.8% ± 10.5%	46.9% ± 9.6%
RF	41.7% ± 7.9%	82.8% ± 6.4%	71.9% ± 7.6%	65.1% ± 4.5%	62.3% ± 5.2%
MLP	**66.1% ± 5.8%**	69.6% ± 14.4%	74.0% ± 7.1%	68.1% ± 9.7%	67.8% ± 9.1%
LDA	44.5% ± 11.6%	76.7% ± 12.9%	69.7% ± 11.5%	62.8% ± 8.6%	60.6% ± 8.7%
KNN	28.0% ± 16.6%	**85.7% ± 9.6%**	63.3% ± 8.1%	61.0% ± 5.4%	56.9% ± 6.7%

Random Forest (RF) also performed well in terms of specificity (82.8% ±6.4%) and AUROC (71.9% ±7.6%), indicating strong discrimination capabilities, though its relatively low sensitivity (41.7% ±7.9%) implies under-identification of ASD cases. More importantly, the substantial increase in performance when going from Decision Tree (DT) to RF for all metrics except sensitivity suggests that our efforts to reduce variance have been effective.

SVM and LDA yielded moderate results, with LDA offering better specificity and AUROC, while SVM had slightly higher sensitivity, yet both classifiers obtained quite low sensitivity rates. DT and KNN performed the worst overall, with the lowest balanced accuracies and high variability. Despite KNN reaching the highest specificity (85.7% ±9.6%), it exhibited the lowest sensitivity (28.0% ±16.6%), reflecting a strong bias toward the control class.

These results indicate that LR and MLP offer the best overall performance on this dataset, with LR slightly outperforming others in consistency and balanced classification, while MLP provides improved ASD detection at the cost of increased false positives.

4.1.2 Performance on Female Data

The performance results of our classifiers on the female-only subset (10 ASD vs. 26 TC) in Table 4.2 reveal substantial challenges in achieving consistent ASD detection for this demographic. Due to the imbalance of female samples between the two classes, a strong bias toward the control class was expected. Although several models achieved perfect specificity (100.0%), this was often accompanied by very low or zero sensitivity, suggesting indeed a tendency to classify all female subjects as controls. LR, SVM, and MLP were the only models to achieve nonzero sensitivity. Among them, MLP achieved the highest sensitivity (60.0% ±41.8%) and a strong AUROC (86.0%± 15.2%), suggesting relatively balanced performance.

The DT, SVM, and MLP achieved a better sensitivity than the dummy classifier, whereas LR scored exactly the same. Regarding the specificity, DT and MLP performed worse than the baseline. For the AUROC and accuracy all models, except DT, performed better than the baseline. However, when it comes to balanced accuracy, most models perform worse, with the exception of LR, SVM, and MLP.

Table 4.2: The average performance across folds on the female data from the NYU site using Pearson correlation features.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
Dummy	30.0% ± 27.4%	84.7% ± 16.6%	50.7% ± 16.0%	69.6% ± 10.9%	57.3% ± 13.0%.
LR	30.0% ± 27.4%	100.0% ± 0.0%	**86.7% ± 11.1%**	80.4% ± 8.2%	65.0% ± 13.7%
SVM	50.0% ± 50.0%	100.0% ± 0.0%	83.7% ± 18.7%	**86.4% ± 13.5%**	**75.0% ± 25.0%**
DT	40.0% ± 22.4%	60.7% ± 32.2%	50.3% ± 19.7%	55.0% ± 24.2%	50.3% ± 19.7%
RF	0.0% ± 0.0%	100.0% ± 0.0%	78.2% ± 25.8%	72.1% ± 1.6%	50.0% ± 0.0%
MLP	**60.0% ± 41.8%**	77.3% ± 15.3%	86.0% ± 15.2%	72.1% ± 1.6%	68.7% ± 13.4%
LDA	0.0% ± 0.0%	100.0% ± 0.0%	82.3% ± 16.7%	72.1% ± 1.6%	50.0% ± 0.0%
KNN	0.0% ± 0.0%	100.0% ± 0.0%	65.3% ± 18.2%	72.1% ± 1.6%	50.0% ± 0.0%

SVM achieved the best overall accuracy (86.4% ±13.5%) and balanced accuracy (75.0% ±25.0%), despite high variability. LR followed closely with a high AUROC (86.7% ±11.1%) and accuracy (80.4% ±8.2%), but a lower sensitivity (30.0% ±27.4%) and lower variability. Random Forest, LDA, and KNN all achieved perfect specificity but failed to identify any ASD case, resulting in a fixed balanced accuracy of 50.0%, which means that all test samples were classified as control. These biases might be caused by a too big bias trade-off for the Random Forest and to a k of 5 for the KNN with only 10 samples in our minority class and could possibly be resolved by reducing said k parameter. It's also important to note that each of RF, LDA and KNN did not use hyperparameter tuning because of non-functional requirement **NFR7**, which might highlight the importance of hyperparameter tuning for under-represented groups despite longer computational times.

These results highlight both the difficulty of training reliable classifiers on the relatively small female subset and a systemic bias toward the majority (control) class in several models. MLP and SVM show potential, but high standard deviations point to instability, likely caused by limited sample size and class imbalance.

4.1.3 Performance on Male Data

The results on the male-only subset (64 ASD vs. 72 TC) in Table 4.3 show a more balanced classifier performance compared to the female subset, possibly due to the larger and more evenly distributed sample size. Among the models, LR achieved the highest AUROC (71.1% ±9.5%) and overall accuracy (63.3% ±5.0%), while MLP yielded the best sensitivity (67.2% ±8.4%) and balanced accuracy (63.4% ±7.0%), indicating a slightly stronger ability to identify ASD cases. This makes MLP a strong candidate for male-specific classification when prioritising recall.

As a baseline, the dummy classifier achieved a sensitivity of 40.6%±8.4%, specificity of 55.5%±8.1%, AUROC of 43.9%±15.8%, accuracy of 48.6%±7.4%, and balanced accuracy of 48.1%±7.4%. In comparison, all models outperformed this baseline.

Table 4.3: The average performance across folds on the male data from the NYU site using Pearson correlation features.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	57.8% ± 19.2%	67.7% ± 12.5%	**71.1% ± 9.5%**	**63.3% ± 5.0%**	62.8% ± 5.4%
SVM	54.9% ± 10.7%	63.9% ± 11.2%	64.4% ± 4.4%	59.6% ± 4.3%	59.4% ± 4.3%
DT	51.5% ± 10.0%	59.6% ± 9.7%	58.4% ± 7.3%	55.9% ± 5.0%	55.6% ± 4.9%
RF	47.1% ± 16.1%	**77.9% ± 6.9%**	69.1% ± 7.4%	63.3% ± 7.7%	62.5% ± 8.3%
MLP	**67.2% ± 8.4%**	59.7% ± 9.9%	65.4% ± 7.3%	63.2% ± 7.2%	**63.4% ± 7.0%**
LDA	51.5% ± 10.0%	69.3% ± 8.3%	63.1% ± 8.2%	61.0% ± 6.9%	60.4% ± 6.9%
KNN	37.2% ± 14.3%	73.7% ± 12.6%	58.6% ± 9.4%	56.7% ± 9.1%	55.4% ± 9.5%

Random Forest achieved the highest specificity (77.9% ±6.9%) and also demonstrated good AUROC (69.1% ±7.4%), though its sensitivity (47.1% ±16.1%) remained relatively low, indicating a more conservative bias toward control predictions. SVM and LDA showed relatively similar and moderate performance across all metrics, with slightly better balance in LDA's specificity-sensitivity trade-off. Decision Tree and KNN were again the lowest-performing models in terms of AUROC and balanced accuracy. In particular, KNN exhibited the lowest sensitivity (37.2% ±14.3%), despite achieving reasonably high specificity, reinforcing its tendency to misclassify ASD cases in this context.

Overall, LR and MLP performed most consistently across metrics, with MLP offering improved sensitivity at a minor cost to specificity. RF was more conservative but discriminative, and the remaining models showed varying degrees of imbalance or instability.

These results also show that the male data did not seem to benefit as significantly from being separately assessed as the female data did, except for reducing the variance a bit, quite possibly due to a more uniform, yet different, connectivity pattern within female brains diagnosed with ASD compared to male brains. Although it does not seem to benefit from this separation, all models outperform the baseline suggesting that the models have learned meaningful patterns from the data and thus has predictive value, whereas that is not always the case for the models on female-only assessment.

4.1.4 Evaluation per sex

Given known differences in the expression and detection of ASD between sexes [38], [39], an additional analysis was performed to evaluate the performance per sex of the the classifiers trained on the combined data. Table 4.4 presents the results, offering insight into potential model biases or differential sensitivity.

Table 4.4: The average performance across folds on the combined data from the NYU site using Pearson correlation features, evaluated per sex.

	Female					**Male**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	**63.3 ± 41.5**	95.0±11.2	**89.4±16.1**	**86.2±13.5**	**79.2±21.2**	58.1±17.0	76.6±6.7	68.6±6.6	**68.0±6.0**	**67.3±6.5**
SVM	36.7 ± 41.5	75.0±25.0	57.7±29.9	63.4±29.5	55.8±27.3	51.4±14.7	65.3±21.4	60.4±13.7	58.3±14.3	58.4±14.2
DT	33.3 ± 47.1	57.7±16.3	48.0±32.8	50.1±24.2	45.5±29.6	42.0±7.6	49.0±21.1	46.7±15.2	46.1±12.3	45.5±11.9
RF	20.0 ± 27.4	93.2±10.9	70.2±38.3	68.0±20.4	56.6±15.7	45.3±9.8	80.3±5.5	**71.0±3.9**	63.2±3.1	62.8±4.6
MLP	56.7 ± 43.5	89.5±14.3	87.7±17.3	78.7±20.5	73.1±28.1	**68.8±7.7**	64.5±14.9	69.6±5.8	66.6±8.1	66.6±7.4
LDA	56.7 ± 25.3	**96.4±8.1**	79.7±21.7	82.2±11.6	76.5±13.8	44.0±15.0	72.4±14.2	66.2±9.5	58.9±8.2	58.2±9.7
KNN	36.7 ± 41.5	**96.4±8.1**	60.8±38.6	75.6±19.0	66.5±23.0	27.8±16.2	**83.8±11.2**	63.9±6.9	57.2±6.5	55.8±7.0

Abbreviations: SEN: sensitivity; SPE: specificity; AUROC: Area Under the Receiver Operating Characteristic Curve; ACC: accuracy; BACC: balanced accuracy.

The results demonstrate notable differences in classifier behaviour when evaluated separately by sex, despite being trained on the same combined dataset. Overall, models tended to perform better on the female subgroup in terms of specificity and AUROC, but with considerably higher variability, especially in sensitivity. This reflects both the limited number of female ASD samples and possibly distinct feature patterns by sex.

LR, LDA, and MLP consistently performed well on female subjects. LR achieved the highest AUROC and accuracy, while it obtained a high specificity and sensitivity, resulting in the high balanced accuracy. MLP offered a strong balanced across both sexes with 73.1% ±28.1% and 66.6% ±7.4% balanced accuracy with a slightly drop in sensitivity (56.7% ±43.5%) for women. However, standard deviations were considerably higher in female metrics, indicating unstable predictions likely caused by the minimal female ASD representation.

In contrast, model performance on male data was more stable but generally lower in AUROC and specificity. MLP achieved the highest sensitivity (68.8% ±7.7%) and second best balanced accuracy (66.6%±7.4%) in males, while LR performed similarly except in sensitivity (58.1%±17.0%). Random Forest showed strong specificity (80.3%±5.5%) but low sensitivity (45.3%± 9.8%), indicating a conservative classification tendency. Decision Tree and KNN underperformed across both sexes, particularly in balanced accuracy.

These findings suggest that while some classifiers (LR and MLP) generalized well across sexes, models struggled more with the female subgroup, possibly due to sample scarcity, leading to greater performance variations.

4.1.5 Evaluation per age

Table 4.5 reports the best-performing and worst-performing age groups. The 30+ age group is excluded in the results, as there were only 3 subjects older than 30 in the test set.

Table 4.5: The average performance across folds on the combined data from the NYU site using Pearson correlation features, evaluated per age group. Only best (left) and worst (right) performing age groups are shown.

	0-11 years					**19-30 years**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	81.0 ± 21.1	69.3±10.9	77.4±14.4	**74.5±12.9**	**75.1±12.7**	42.3±32.7	**97.5±5.9**	**71.2±27.9**	**69.5±21.4**	**69.9±17.0**
SVM	74.3 ± 25.1	71.4±8.0	72.1±15.2	70.5±12.3	72.9±10.7	29.0±41.3	78.3±33.1	57.3±32.9	53.3±30.5	53.7±32.9
DT	42.4 ± 17.6	46.4±33.0	45.1±19.2	43.9±19.5	44.4±20.1	**58.0±26.6**	69.3±23.0	59.3±16.1	57.7±11.0	63.7±10.4
RF	64.9 ± 25.0	**74.3±1.6**	**78.8±9.5**	66.6±13.1	69.6±11.9	10.7±15.3	**97.5±5.6**	70.8±18.6	62.6±20.2	54.1±9.1
MLP	**86.7 ± 21.7**	56.4±20.7	77.3±12.9	71.9±9.7	71.5±8.4	34.0±25.9	82.5±24.4	67.3±19.2	65.5±22.6	58.3±15.8
LDA	74.8 ± 23.8	66.4±12.0	77.4±10.4	68.5±7.6	70.6±6.5	5.0±11.2	88.5±16.9	65.3±25.2	54.0±22.2	46.75±11.8
KNN	44.9 ± 29.3	70.0±32.6	66.5±21.1	55.2±23.1	57.5±19.4	0.0±0.0	**97.5±5.6**	55.8±18.7	56.9±20.7	48.9±2.8

Overall, all classifiers exhibited higher sensitivity, balanced accuracy, and AUROC in the 0–11 group than in the 19–30 group, suggesting more distinguishable patterns between ASD and control samples in younger subjects. On an important note, the NYU dataset contains 53 samples in the 0–11 age group, 72 samples in the 12–18 group, 42 samples in the 19–30 age group, and 5 samples in the 30+ group.

LR, SVM, MLP, and LDA showed particularly strong performance in the 0–11 group. MLP achieved the highest sensitivity and accuracy, while LR produced the best balanced accuracy with a good AUROC. LDA also performed well, reaching a balanced accuracy of 70.6% ± 6.5% with relatively low variability. In contrast, performance across all classifiers, except DT, degraded in the 19–30 group. Most notably, sensitivity dropped substantially for LDA (5.0% ± 11.2%) and RF (10.7% ± 15.3%), while KNN failed to detect any ASD cases.

This decline in sensitivity and balanced accuracy in older subjects may reflect increased heterogeneity in ASD presentation with age or reduced model generalisation to adult brain connectivity patterns. These findings underscore the potential benefit of age-specific modelling strategies, such as multitask learning to account for the age differences, particularly for adult populations where classifier performance was consistently less reliable.

4.2 Discussion

The evaluation of classifier performance on the NYU subset of the ABIDE I dataset highlights several important trends and challenges in ASD classification using functional connectivity features.

4.2.1 Classifier Behavior Across Subsets

Overall, LR and MLP emerged as the most robust classifiers across all demographic settings. LR consistently yielded high AUROC and balanced accuracy scores, demonstrating its ability to generalize well despite variations in the data. MLP, on the other hand, exhibited stronger sensitivity, particularly in the combined and male subsets, indicating greater potential for ASD detection at the cost of some false positives.
The Decision Tree (DT) and KNN models underperformed in nearly all settings. With KNN displaying strong biases toward the majority (control) class. This may be attributed to an unoptimized choice of k relative to the minority class sizes. Random Forest (RF) improved upon DT as expected, showing strong specificity and AUROC, validating its design to reduce variance via ensemble learning. However, its sensitivity remained relatively low, especially in sex- and age-specific evaluations, possibly due to conservative decision boundaries. On the contrary, DT performed better on young adults, while all other models displayed a strong decrease in performance.

4.2.2 Sex-Specific Observations

Performance on the female subset was significantly less stable compared to the male subset, with large standard deviations and multiple models achieving zero sensitivity. This behaviour indicates a systemic bias where models tend to classify all female subjects as controls. However, the female samples showed much better performance for both the specialised and generalised model compared to the males. Here, the specialised model refers to a sex-specific model trained exclusively on data from one sex (i.e., separate models trained on female-only and male-only data), while the generalised model is trained on the combined dataset (female and male samples together) and evaluated separately for each sex. Contributing factors likely include the limited number of female ASD samples and potential differences in functional connectivity patterns between sexes. While MLP and SVM showed potential for female specialised models, the variance suggests these models remain sensitive to the imbalance and are prone to instability. This is likely the result of our small and unbalanced female dataset.
Evaluations on the male subset showed more consistent results across classifiers, likely due to the more balanced representation of ASD and control samples. Both LR and MLP again demonstrated strong performance, reinforcing their relative robustness. However, gains from analysing males separately were marginal compared to females, implying that performance disparities stem primarily from sample scarcity rather than fundamental differences in male brain connectivity patterns.

4.2.3 Age-Related Trends

Classifier performance was highest for the youngest age group (0–11 years) and lowest for the oldest (19–30 years), particularly in terms of sensitivity and balanced accuracy. Models such as LR, MLP, and LDA achieved strong performance in the youngest group, while many classifiers, including RF, LDA, and KNN, struggled to detect any ASD in the older group. This drop may be due to increased heterogeneity in brain connectivity patterns with age or due to insufficient representation of adult samples, reducing generalisation capabilities.

4.2.4 Implications and Limitations

The findings highlight that ASD detection performance is highly sensitive to demographic characteristics, particularly sex and age. This suggests the potential need for demographic-specific models or preprocessing strategies, such as balancing, resampling, or domain adaptation. Furthermore, several classifiers, especially those not tuned, suffered disproportionately in under-represented groups, emphasising the importance of hyperparameter tuning when data is limited or imbalanced.
The use of only Pearson correlation features from fMRI also represents a simplification, and future work could explore whether partial correlations, dynamic connectivity measures, or multimodal integration (e.g., structural MRI or phenotypic data) would yield more robust models. Additionally, though this chapter focused on a single acquisition site to reduce inter-site variability, this limits the generalisability of results to other populations and scanner protocols.
Furthermore, it is important to consider that the NYU dataset contains a relatively low number of subjects (172) in comparison to the 6670 Pearson correlation features used, with 36 female subjects and 136 male subjects. Training on such low numbers will highly likely risk overfitting of our classifiers. In addition, since the number of samples is sparse in a single-site dataset, to maximise the training data it may be more appropriate

to use 10-fold cross validation instead of 5-fold cross validation. For evaluating the female subset it may be even better if analysis if done with Leave-One-Out cross validation.

4.2.5 Future Directions

To improve generalisation and fairness, future work could explore data augmentation or synthetic oversampling for underrepresented subgroups, fine-grained hyperparameter tuning tailored to demographics, and domain-aware feature selection strategies. Furthermore, implementing Leave-One-Out CV may offer new insights, as this maximises the training data, which is highly preferable when sample size is sparse. Additionally, we suggest investigating the use of Histogram Gradient Boosting Trees (HGBT) as an alternative to Random Forest, Quadratic Discriminant Analysis (QDA) as a counterpart to LDA and the possibilities of AdaBoost, an ensemble learning technique that initially assigns equal weights to all training samples and iteratively adjusts them to focus on misclassified instances. HGBT offers higher computational speed for both fit and prediction with lower likelihood to overfit. This increased speed could also allow for hyperparameter tuning without failing to comply to **NFR7**, demanding a maximum model computation time of 40 minutes. Introducing QDA could possibly increase sensitivity and mitigate the bias induced by LDA within minority groups. AdaBoost would be a good addition to analyse due to its interpretable nature, low likelihood to overfit, and great handling of unbalanced data [40].

In the next chapter, Chapter 5, we will continue with validation on multi-site data with harmonisation techniques which may also help verify whether observed patterns persist outside the NYU cohort. In Chapter 6 we will demonstrate how developing interpretable models could aid in understanding the specific brain regions or connections driving classification decisions.

5 Multi-Site Performance on Pearson Correlation

In the following, we present our results in the multisite classification setting on Pearson correlation features.

5.1 Non-Harmonised versus Harmonised Classification

Table 5.1 shows the results of 5-fold cross validation for multisite classification on non-harmonised (raw) data and multisite classification on the full harmonised dataset.
All classification models outperform the dummy baseline (DUM) on all metrics, except for RF on sensitivity which is below the baseline (-4.6%). LR, SVM, MLP, and LDA perform really well on the raw data adhering to the set of requirements (**NFR1**, **NFR2**, **NFR3**, **NFR4**, **NFR5**). DT and KNN do not conform to these requirements.
On the harmonised data, again RF underperforms on sensitivity. For all other metrics, all classifier models outperform the baseline.

Table 5.1: The average performance across folds on the raw combined multi-site data versus the harmonised combined multi-site data using Pearson correlation features using Pearson correlation features.

	Raw					**Harmonised**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
DUM	48.6 ± 4.1	54.9±5.2	47.8±2.4	52.0±4.1	51.7±4.1	48.4±5.5	54.0±6.5	51.7±3.8	51.4±2.9	51.2±2.8
LR	58.7 ± 2.7	66.1±2.8	67.8±3.2	62.7±1.8	62.4±1.8	60.2±4.9	66.1±2.4	68.1±3.4	63.4±2.4	63.1±2.5
SVM	59.8 ± 4.1	69.0±4.7	68.3±1.5	**64.7±2.3**	**64.4±2.2**	55.2±6.5	65.6±4.9	64.8±5.5	60.8±4.0	60.4±4.0
DT	54.4 ± 6.7	58.1±7.0	57.4±4.0	56.4±3.4	57.0±5.2	49.2±19.0	59.7±15.2	56.9±2.6	54.8±2.1	54.4±2.8
RF	42.0 ± 6.5	**80.1±3.1**	68.1±4.4	62.6±2.9	61.1±3.2	39.8±5.5	**76.5±8.7**	64.5±5.6	59.5±6.3	58.1±6.2
MLP	**61.7 ± 4.4**	67.0±5.0	**69.6±2.8**	64.5±2.5	64.3±2.5	**63.2±4.7**	65.0±6.5	**69.8±3.6**	**64.2±4.9**	**64.1±4.9**
LDA	58.0 ± 5.5	68.5±3.6	68.2±2.9	63.6±3.1	63.2±3.2	53.9±6.3	62.0±5.6	59.8±6.5	58.3±5.3	57.9±5.3
KNN	56.2 ± 9.0	61.6±4.3	62.2±4.1	59.1±2.7	58.9±3.1	58.5±10.4	59.8±4.6	62.0±4.1	59.2±3.7	59.2±4.1

Comparing the results of our classifiers on the raw data versus the harmonised data, it can be seen that for most models the performance decreases when harmonisation is applied. Only LR and KNN seem to benefit from harmonisation. For LR, all performance metrics increase when harmonisation is applied with the exception of specificity. However, the specificity of LR is a bit more stable for the harmonised data than for the raw data (-0.4%). In other models, harmonisation seems to make the performance more unstable with the standard deviation of the sensitivity of the DT model going as high as 19.0%. Overall, it seems that the application of harmonisation introduces instability in results, and the increase in performance is very minimal.

5.2 Evaluation per site

Given the limited advantages of harmonising the features before classification, the following results have been obtained using the raw, non-harmonised, data. Table 5.2 presents the sites that performed best and worst in terms of balanced accuracy during the per-site evaluation. The per-site performance for all individual sites can be found in Section F.1.

Analysing the performance of classifiers across the 20 individual ABIDE I sites revealed substantial variability. While some sites consistently supported accurate classification, others showed high sensitivity to model choice and exhibit large performance inconsistencies.

Table 5.2: The average performance across folds on raw combined data using Pearson correlation features, evaluated per site. Only best (left: PITT) and worst (right: OHSU) performing sites are shown. PITT contained 24 ASD and 26 TC samples in total. OHSU contained 12 ASD and 13 TC samples in total.

	PITT					**OHSU**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
DUM	42.8±13.0	57.5±13.2	39.1±11.9	51.8±8.8	50.1±9.3	41.7±34.7	54.2±41.1	39.1±11.9	51.8±8.8	50.1±9.3
LR	63.9±29.9	**91.0±12.4**	**82.0±20.0**	**76.5±18.2**	**77.4±17.4**	37.5±28.5	20.8±25.0	32.3±37.3	32.1±23.7	29.2±25.0
SVM	58.9±16.0	84.5±15.9	77.1±18.5	71.4±12.5	71.7±12.2	25.0±31.9	27.1±35.6	23.3±27.9	32.1±23.7	26.0±22.1
DT	59.4±38.6	68.0±12.5	61.3±29.2	60.6±21.9	63.7±23.2	**50.0±40.8**	47.9±44.3	**49.7±40.5**	43.5±35.4	**49.0±38.2**
RF	43.3±36.5	87.5±12.5	66.9±23.0	65.5±23.1	65.4±22.4	25.0±21.5	33.3±23.6	38.2±20.1	32.7±14.6	29.2±17.3
MLP	58.3±30.0	90.0±13.7	81.1±16.8	72.5±16.6	74.2±14.8	33.3±30.4	31.2±37.5	39.6±36.9	36.9±21.8	32.3±22.7
LDA	60.6±12.8	86.0±12.9	73.5±4.8	72.7±5.7	73.3±6.7	33.3±30.4	**50.0±40.8**	26.0±32.3	**45.2±29.3**	41.7±32.6
KNN	**67.2±27.7**	72.0±12.5	69.9±18.2	70.3±10.9	69.6±13.7	37.5±47.9	35.4±29.2	36.8±37.5	44.0±28.3	36.5±31.8

Across classifiers, LDA and LR demonstrate relatively stable performance, especially at sites like NYU, USM, and YALE.
Most classifiers outperformed the baseline by a great margin. The RF model barely outperforms the baseline on sensitivity by a small margin of +0.7%. LR obtained the highest balanced accuracy, going as high as 77.4%.
In contrast, the performance of our classifiers on OHSU were very low. Only DT managed to outperform the baseline by a relatively low margin of 1.1%. On an important note, all classifiers, as well as the baseline, displayed highly unstable performance with standard deviations sometimes reaching higher than 40%.
Notably, almost all models perform very poorly on Caltech with balanced accuracies of below 50%, but RF scores very high on this site, with a balanced accuracy of 91.7% (±14.4%). In addition, all models performed moderately to poorly on STANFORD, with the highest balanced accuracy being 55.0% by the LDA model.

5.3 Evaluation per sex

Table 5.3 presents the performance of the seven classifiers on combined raw multi-site data, evaluated per sex. Remarkably, the performance of the classifiers is differently affected by the new data for male and female analysis. Interestingly, models that initially performed poorly on female data during per-sex evaluation on combined NYU data (SVM, DT, RF), show strong improvement on the multisite data, while models that previously performed well in this setting (LR, LDA, MLP) exhibit moderate drops.
All models outperformed the baseline set by the dummy classifier on female data in terms of balanced accuracy, with the lowest margin being +2.1% for our LDA model. Some models perform worse than the baseline when it comes to sensitivity. Only MLP outperforms by a great margin on this metric. As for AUROC and specificity, all models outperform the baseline. All in all, it seems that MLP performs the best on the female data. RF and LR are also strong contenders.
All models outperform the baseline on the male data set by the dummy classifier by a great margin, with the only exception of the sensitivity obtained by our RF model (-6.3%).

Table 5.3: The average performance across folds on raw combined data of all sites using Pearson correlation features, evaluated per sex.

	Female					**Male**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
DUM	48.9±12.5	56.3±11.2	44.9±9.6	53.7±8.4	52.6±8.5	48.5±2.9	54.4±4.4	48.1±3.9	51.6±3.4	51.4±3.3
LR	46.1±24.7	83.0±3.0	71.4±18.1	69.7±9.6	64.6±13.1	60.6±3.2	62.1±2.8	66.7±2.6	61.4±1.6	61.4±1.6
SVM	50.5±28.6	75.0±16.0	69.5±18.8	67.4±7.3	62.7±9.7	59.1±2.4	63.7±3.8	65.0±2.7	61.5±1.1	61.4±1.0
DT	54.4±17.7	55.0±7.5	58.0±15.4	54.7±10.2	54.7±11.9	54.4±5.3	58.5±9.5	56.9±3.0	56.6±3.7	56.5±3.6
RF	38.1±16.6	**92.2±6.2**	**74.7±8.8**	72.6±8.1	65.2±9.2	42.6±6.1	**77.4±3.0**	66.6±4.3	60.7±3.0	60.0±3.0
MLP	**62.8±28.6**	78.7±10.9	71.9±16.9	**73.2±14.3**	**70.7±16.6**	**61.7±3.6**	64.1±4.8	**68.9±1.8**	62.9±1.7	62.9±1.6
LDA	46.1±14.8	75.1±12.4	69.5±13.2	65.1±10.5	60.6±10.5	59.6±5.4	66.9±4.6	67.5±1.9	**63.4±2.4**	**63.3±2.4**
KNN	52.6±20.6	70.2±9.4	68.0±14.5	63.5±10.4	61.4±11.4	57.1±8.6	59.5±3.2	61.0±4.1	58.3±3.0	58.3±3.1

Comparing the two, LR, SVM, RF, MLP, and KNN perform better on the female subset, than on the male subset in terms of balanced accuracy. This difference is largest in the balanced accuracy achieved by MLP, namely a difference of 7.8%. For the other four models, this difference is not as large and usually not more than a 3% difference is seen. Only DT and LDA perform better on male data than female data.

5.4 Evaluation per age

Table 5.4 compares classifier performance on the raw combined multi-site dataset across two age groups: adolescents (12–18 years), which represent the best performing group, and children (0–11 years), the worst performing group, in terms of balanced accuracy. The complete results of all age group can be found in Table F.11 and F.12. The metrics reveal a consistent trend of superior classification performance for adolescents across nearly all models and evaluation criteria.

Table 5.4: The average performance across folds on raw combined data using Pearson correlation features, evaluated per age group . Comparison between **12-18** (left) and **0-11** (right).

	12-18					**0-11**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
DUM	45.2±9.9	54.5±7.8	46.5±3.3	50.1±7.6	49.9±7.7	58.3±14.1	55.4±12.4	46.5±11.5	57.0±12.6	56.9%±12.6
LR	61.1±6.8	69.2±8.6	72.9±3.6	65.0±1.2	65.1±1.1	57.8±13.3	55.9±9.3	61.7±10.1	55.8±8.4	56.8±9.3
SVM	61.9±7.0	69.1±7.5	71.2±3.8	65.5±1.0	65.5±1.2	52.4±6.0	57.7±9.8	59.8±8.7	55.2±7.9	55.0±7.8
DT	49.6±11.4	60.1±6.6	55.9±6.6	55.3±4.7	54.8±4.4	59.4±22.7	54.7±9.2	60.7±15.2	56.7±13.3	57.0±13.2
RF	44.0±11.9	**82.4±3.1**	70.7±6.4	64.4±5.7	63.2±6.0	42.9±10.8	**66.4±7.7**	59.5±10.2	56.3±5.0	54.7±5.6
MLP	**65.2±6.4**	69.1±8.4	**73.5±4.0**	**66.9±4.1**	**67.1±3.9**	**61.9±14.9**	54.4±9.0	**64.3±10.4**	57.0±7.2	58.2±6.9
LDA	60.5±7.6	68.6±6.1	70.0±4.8	64.6±5.8	64.5±6.0	61.3±13.5	58.7±9.9	63.6±12.6	**59.2±8.4**	**60.0±8.2**
KNN	60.3±10.2	66.7±5.6	67.3±5.4	63.3±4.2	63.5±4.3	52.7±14.2	57.0±6.1	58.6±8.9	53.6±5.2	54.8±5.5

In the adolescent group, classifiers show balanced and relatively high sensitivity and specificity scores. MLP achieves the highest AUROC (73.5% ± 4.0), alongside the best sensitivity (65.2% ± 6.4), closely followed by LR and SVM. RF also shows strong specificity (82.4% ± 3.1), which contributes to its solid accuracy (64.4% ± 5.7) despite a lower sensitivity. Overall, balanced accuracy for most classifiers in this group lies between 64–66%, with the exception of DT (54.8% ± 4.4), indicating reliable ASD recognition potential.

In contrast, performance in the child group (0–11 years) is generally weaker and more inconsistent, as the standard deviation is quite high. The accuracy and balanced accuracy performance of LR on children decreased significantly with -9.2%, and -7.3%, respectively. Almost all classifier models show a sharp decrease in balanced accuracy, except for DT, who actually obtained a higher balanced accuracy in children. However, the standard deviation for this metric is very high (13.2%), so its performance is not reliable.
These findings suggest that age significantly affects classification efficacy, likely due to developmental differences in brain connectivity patterns and the increased noise or motion artifacts in younger subjects' fMRI data. Furthermore, younger participants are possibly underrepresented or unevenly distributed across sites, which may exacerbate the model's challenges in learning reliable age-specific features.

Only DT, MLP, and LDA managed to outperform the baseline in terms of sensitivity. DT outperforms the baseline on sensitivity by a rather small margin (+1.1%), while MLP and LDA outperform the baseline more confidently. LR manages to come close to the sensitivity of the baseline, but falls just short (-0.5%). In terms of accuracy, only LDA outperforms the baseline. MLP achieved the same accuracy, and was more stable. Lastly, comparing balanced accuracy results, only DT (+0.1%), MLP (+1.3%), and LDA (+3.1%) outperform the baseline. However, it should be noted that the performance is much more stable.
Future work may consider age-stratified preprocessing pipelines, age-adaptive models, or targeted feature selection to improve classification robustness for varying age groups. The overall patterns underscore the importance of demographic segmentation when evaluating ASD classification performance on multisite neuroimaging data.

5.5 Cross-site Generalisation with LOGO CV

Table 5.5 presents the balanced accuracies obtained by each classifier for each test site on the raw combined multisite data. Details of the other performance metrics are shown in Appendix F. The results in Table 5.5 shows considerable variation in balanced accuracy across test sites and classifiers, highlighting the difficulty of generalising to unseen imaging sites. On average, the MLP and SVM classifiers achieved the highest mean balanced accuracy (63.9% and 61.8%, respectively), followed closely by Logistic Regression (61.3%). In contrast, DT, KNN, and LDA consistently underperformed, with means of 57.5%, 56.2%, and 59.7%, respectively.

Table 5.5: Balanced accuracy obtained by each classifier for each test site during LOGO CV on Pearson correlation features. N indicates the number of ASD and TC samples in each site, which also corresponds to the size of the test set for that fold in LOGO CV, where each group (site) is left out once for testing.

	Balanced Accuracy [%]							**N**	
	LR	SVM	DT	RF	MLP	LDA	KNN	ASD	TC
CALTECH	45.0	45.0	45.0	55.0	**65.0**	40.0	40.0	5	10
CMU	81.7	81.7	71.7	66.7	**91.7**	75.0	38.3	6	5
KKI	**66.1**	63.7	63.7	61.3	**66.1**	**66.1**	**66.1**	12	21
LEUVEN_1	53.6	50.0	50.0	53.6	53.6	**57.1**	**57.1**	14	14
LEUVEN_2	51.0	52.1	57.3	55.2	50.0	56.2	**74.0**	12	16
MAXMUN	48.8	43.6	55.8	57.0	48.1	**60.4**	47.3	19	27
NYU	**64.5**	53.6	58.6	63.2	59.4	59.6	62.3	74	98
OHSU	44.2	35.9	56.4	**63.8**	52.2	39.1	43.9	12	13
OLIN	67.9	67.9	57.1	64.3	**71.4**	67.9	**71.4**	14	14
PITT	**73.2**	69.2	57.9	52.7	**73.2**	65.1	63.8	24	26
SBL	35.7	**59.5**	35.1	58.9	47.6	55.4	41.7	12	14
SDSU	**73.4**	68.1	45.1	61.8	64.5	60.2	60.2	8	19
STANFORD	56.4	56.4	63.8	**67.9**	56.4	48.7	36.9	12	13
TRINITY	56.2	**72.8**	56.9	49.7	62.8	55.7	55.6	19	25
UCLA_1	**72.3**	69.6	60.5	57.5	67.8	63.1	68.3	37	27
UCLA_2	70.9	70.9	**85.9**	66.8	65.9	70.5	66.8	11	10
UM_1	57.6	61.5	**69.9**	63.1	58.7	59.3	64.2	34	52
UM_2	62.6	68.3	52.0	65.9	**72.2**	64.5	49.3	13	21
USM	71.5	73.5	60.8	64.9	**75.6**	68.2	55.2	43	24
YALE	73.6	73.6	47.1	59.6	**75.8**	62.6	61.8	22	19
Mean±std	61.3±11.9	61.8±11.7	57.5±10.7	60.4±5.2	63.9±10.8	59.7±8.9	56.2±11.3		

Some sites, such as CMU, OLIN, and YALE, showed relatively high generalisation performance across classifiers. The balanced accuracy results on MAXMUN and OHSU are remarkedly low.
No single classifier consistently outperformed others across all sites. MLP and SVM tended to perform well across a broader range of test sites, potentially reflecting greater ability to adapt to unseen distribution of sites. Classical linear models, like Logistic Regression also yielded competitive performance, suggesting robustness under distribution shifts. These findings reinforce the challenge of cross-site generalisation in ASD classification. Even when trained on a large, multi-site dataset, models may struggle to generalise to sites with different acquisition characteristics.

5.6 Discussion

This chapter presented a comprehensive evaluation of ASD classification using Pearson correlation rs-fMRI features across multiple acquisition sites from the ABIDE I dataset. Several classifiers and evaluation protocols were compared, including within-site validation, stratified multi-site analysis, and cross-site generalisation using Leave-One-Group-Out (LOGO) cross-validation. The results yielded several insights into the impact of harmonisation, demographics, and inter-site variability on classification performance.

5.6.1 Harmonisation versus Raw Multisite Training

Harmonisation via NeuroHarmonize (based on ComBat) was hypothesised to reduce site-specific biases. However, as shown in Section 5.1, the difference between harmonised and raw data performance was minimal. In most cases, harmonisation actually lead to a performance drop of our classifiers and there was a noticeable increase of instability. We suspect that the harmonisation technique may have overcorrected for the site variability and in doing so actually removed meaningful biological signals important for ASD detection.

5.6.2 Site-Specific Variability

Section 5.2 demonstrated that per-site performance varies drastically. The classifiers obtained very high performance on PITT across all metrics. In constrast, the performance on OHSU was much lower. This may be explained by the different scanning procedures used by the various acquisition sites. Importantly, most acquisition sites asked their subjects to lie still with their eyes closed. In contrast, OHSU participants were scanned with their eyes open. This may have influenced the measured functional connectivity. The low achieved balanced accuracies on STANFORD may be explained by the slice thickness of their scanner hardware as they used a slice thickness of 4.5mm (see Appendix A.2). Another possible reason may be age-related. The

full ABIDE I dataset has a mean age of 16.94 years (±7.58), while the mean age in the STANFORD dataset is equal to 9.99 years ±1.63. This coincides with the low performance accuracy on the 0-11 age group.
These site-based performance disparities were further highlighted in the LOGO cross-validation results, where classifiers trained on all-but-one sites struggled to generalise to unseen sites. Despite robust performance on internal validation, cross-site generalisation remained limited and a key challenge for clinical translation of ASD classification models. Sites such as USM and CMU yielded notably high balanced accuracies. Conversely, classification performance on sites like MAXMUN was markedly low, likely due to the wide variability in the ages of the MAXMUN subjects, ranging from 7 to 58 years. Furthermore, the mean age of the MAXMUN dataset is 26.5±10.63 years, whereas the mean age of the full data set is around 16.94 years. In addition, MAXMUN used various scan procedures across their subjects. Some were asked to lie still and keep their eyes closed, whereas others were instructed to keep their eyes open and look at a picture of a night sky with stars [41]. The use of two different scan procedures within their study as well as the wide age range might explain the low obtained balanced accuracy across all classifiers.

5.6.3 Demographic Effects

Sex- and age-specific evaluations reveal consistent patterns in classification performance. Female samples, despite being evaluated on the same multisite models, exhibited greater sensitivity improvements in poorly performing classifiers (e.g., SVM, DT, RF), while models that previously excelled (LR, LDA) showed slight declines. However, due to overall under-representation of female participants outside the NYU site, performance gains were limited and variance remained high. This highlights the importance of sex-balanced datasets and stratified evaluation.

Age-based comparisons revealed that adolescents (12–18 years) consistently achieved higher and more stable classification performance across nearly all models. In contrast, younger children (0–11 years) exhibited weaker and more variable results. Several neurodevelopmental and methodological factors likely contribute to this disparity. First, younger participants tend to exhibit significantly greater in-scanner head motion, a well-documented confound in pediatric fMRI studies. While the ABIDE I dataset, preprocessed using the CPAC pipeline, incorporates comprehensive motion correction steps, outlined in Appendix A.1, these procedures cannot fully eliminate the impact of motion. Residual motion artifacts persist even below the FD threshold and can introduce systematic biases in functional connectivity estimates [42, 43].
Second, functional connectivity itself undergoes substantial developmental changes across childhood and adolescence. Prior studies have shown that connectivity patterns in ASD are not static across development. Younger individuals may display both over- and under-connectivity, while adolescents more reliably exhibit under-connectivity [44].
In conclusion, as the brain matures, adolescents tend to exhibit more stable and distinguishable connectivity patterns, making it easier for classifiers to identify ASD-related features. In contrast, younger children show greater developmental variability, which makes these patterns harder to detect and learn from.

5.6.4 Classifier Comparison

Across most evaluations, MLP and SVM achieved strong and consistent performance, particularly in cross-site tests (balanced accuracies of 63.9% and 61.8% respectively). Logistic Regression and MLP where both stable and performed consistent across different demographics, making them appealing when inclusivity and consistency are important, with MLP slightly more reliable for small skewed training sets. Decision Trees and KNN were generally outperformed across all settings, showing sensitivity to noise and site variability. Introducing Random Forest as an enhancement to Decision Trees proved successful, with Random Forest outperforming its predecessor across all settings and demonstrating improved robustness to noise and site variability.

5.6.5 Limitations and Future Work

While the study provides valuable insights, several limitations remain. First, the use of Pearson correlation features may overlook more complex temporal or spectral properties of rs-fMRI data. Second, the imbalance in sex and age distributions across sites likely biased classifier training. Third, the LOGO CV setup assumes equal difficulty between sites, which may not be true given population sizes and demographics. Moreover, data augmentation strategies might be implemented to mitigate variability and improve generalisation.

Future work should explore sex- and age-specific model tuning, incorporate more expressive graph- or network-based features, and examine techniques to explicitly address inter-site heterogeneity. One technique worth looking into is the application of domain adaptation techniques as that may better address the site heterogeneity of the multisite data [45].
Furthermore, as stated in our discussion on single-site classification, the same holds for the multisite setting. With 6670 features in comparison to 871 samples, there is still a high risk of overfitting. Highlighting the need for proper feature selection techniques and possibly dimensionality reduction. We expect that with the integration of the feature selection methods of the Feature Selection subgroup, the performance of all developed classifiers will increase.

5.6.6 Conclusion

In conclusion, this study highlights both the potential and the limitations of ASD classification from multi-site rs-fMRI data. While multisite training enables reasonable within-distribution performance, generalising to unseen sites and under-represented subgroups remains a challenge. Addressing these challenges requires continued efforts in dataset balance, feature design, and robust validation strategies to move toward clinically useful ASD detection tools.
Despite these challenges, our LR, SVM, and MLP models satisfied all performance-related requirements (**NFR1-NFR6**) and adhered to the set threshold for the computation time in **NFR7**. LDA satisfied most requirements, but failed to meet the requirements for the standard deviation in **NFR6**.

6 Feature Importance and Brain Visualisation

This chapter explores which brain connectivity features were most influential in the classification of ASD versus TC. By analysing feature importance across folds and models, we identify patterns of stable, high-ranking connections and examine their relevance. We then visualise the most consistently important connections in the brain, bridging model outputs with neuroscientific interpretation and contributing to the broader goal of biomarker discovery in ASD.

6.1 Stability Across Folds and Models

Table 6.1 and 6.2 present a subset of brain connectivity features that most frequently appeared among the top 20 important features ranking across folds for both the LR and SVM models as these conform to all requirements **NFR1-NFR7**. The complete list of all features identified in the top 20 ranking can be found in Table F.17 in Appendix F.

These results highlight the stability and consistency of certain brain connectivity features across both classifiers and validation folds. In particular, the feature `fc_5301_8212`, representing a connection between the left inferior occipital gyrus (Occipital_Inf_L) and the right superior temporal pole (Temporal_Pole_Sup_R), appeared in 9 out of 10 folds (5 for LR and 4 for SVM), suggesting a highly robust and potentially biologically meaningful connection that may play a key role in distinguishing between ASD and TC. Other features, such as `fc_6302_9160` (connecting the right precuneus and vermis 9) and `fc_2002_8201` (linking the right precentral gyrus with the right Heschl's gyrus), also appeared in the majority of folds across both classifiers, suggesting a pattern of shared feature relevance despite the different learning objectives of logistic regression and SVM.

Importantly, by averaging feature weights only over folds where the features appear in the top 20 ranking, separately per classifier, we observe systematic differences in magnitudes of weights between LR and SVM (see Table REF). For example, `fc_5301_8212` has a notably larger negative average weight magnitude in LR (approximately -0.12) compared to SVM (approximately -0.02). Although the average absolute weight magnitudes were larger for LR than for SVM, this difference reflects the distinct optimisation objectives and constraints of each model.

Table 6.1: Average feature weights (± standard deviation) across folds where the feature appeared in the top 20 for each classifier. The number of folds in which each feature appeared in the top 20 ranking is also shown. Only the subset of features that appeared in the top 20 rankings in more than 6 out of 10 folds in total are shown.

	LR		**SVM**	
Feature	Weight (mean±std)	#Folds	Weight (mean±std)	#Folds
fc_5301_8212	-0.1172±0.0123	5	-0.0198 ±0.0016	4
fc_6302_9160	-0.1071±0.0075	4	-0.0179±0.0014	4
fc_2002_8201	-0.1057±0.0070	4	-0.0179 ± 0.0014	3
fc_2211_2312	0.1084±0.0044	3	0.0180 ± 0.0007	3
fc_2332_9021	-0.1225±0.0087	3	-0.0206 ± 0.0015	3
fc_2201_5102	0.1137±0.0114	3	0.0189 ± 0.0022	3

Index to region mapping: 5301: left inferior occipital gyrus; 8212: right middle temporal pole; 6302: right precuneus; 9160: vermis 9; 2002: right precentral gyrus; 8201: left middle temporal gyrus; 2211: left orbital part of the middle frontal gyrus; 2312: right triangular part of the inferior frontal gyrus; 2332: right Rolandic operculum; 9021: left lobule 3 of the cerebellum; 2201: left middle frontal gyrus; 5102: right superior occipital gyrus.

While a number of features appeared in the top 20 for only a single fold, suggesting some degree of cross-fold variability, there was a notable trend: for a given fold, if a feature was ranked among the top 20 by the logistic regression model, it was often also ranked among the top 20 by the SVM model. This indicates a relatively high degree of agreement between models within the same data partition, even though importance rankings may vary across different folds.

By identifying features that are both stable across models and mapped to specific anatomical regions, this analysis contributes directly to the goal of biomarker discovery. These repeatedly high-ranking connections offer promising targets for neuroscientific investigation into the functional architecture of ASD.

Table 6.2: Often occurring top 20 discriminative brain connectivity features selected across folds and models. Each checkmark indicates that the corresponding feature was among the top 20 most important features for a given fold and model in stratified 5-fold cross-validation.

Feature name	LR-F1	LR-F2	LR-F3	LR-F4	LR-F5	SVM-F1	SVM-F2	SVM-F3	SVM-F4	SVM-F5
fc_5301_8212	✓	✓	✓	✓	✓	✓	✓		✓	✓
fc_6302_9160	✓		✓	✓	✓	✓		✓	✓	✓
fc_2002_8201	✓		✓	✓	✓	✓		✓	✓	
fc_2211_2312	✓		✓	✓		✓		✓	✓	
fc_2332_9021		✓	✓	✓			✓	✓	✓	
fc_2201_5102			✓	✓	✓			✓	✓	✓

Index to region mapping: 5301: left inferior occipital gyrus; 8212: right middle temporal pole; 6302: right precuneus; 9160: vermis 9; 2002: right precentral gyrus; 8201: left middle temporal gyrus; 2211: left orbital part of the middle frontal gyrus; 2312: right triangular part of the inferior frontal gyrus; 2332: right Rolandic operculum; 9021: left lobule 3 of the cerebellum; 2201: left middle frontal gyrus; 5102: right superior occipital gyrus.

Figure 6.1 provides a brain-based visualisation of the six most consistently important connectivity features listed in Table 6.2. This figure offers an intuitive understanding of their spatial distribution and potential involvement in ASD-related neural pathways. This visual representation serves as a crucial bridge between statistical model output and neuroscientific interpretation, highlighting candidate connections for future biomarker validation. Additional connectome plots for each fold and model, showing the top 20 features weighted by their importance scores, are available in Appendix E.1 (Figures E.1-E.10). In Chapter 9, we link the most important features found during this analysis to existing ASD research.

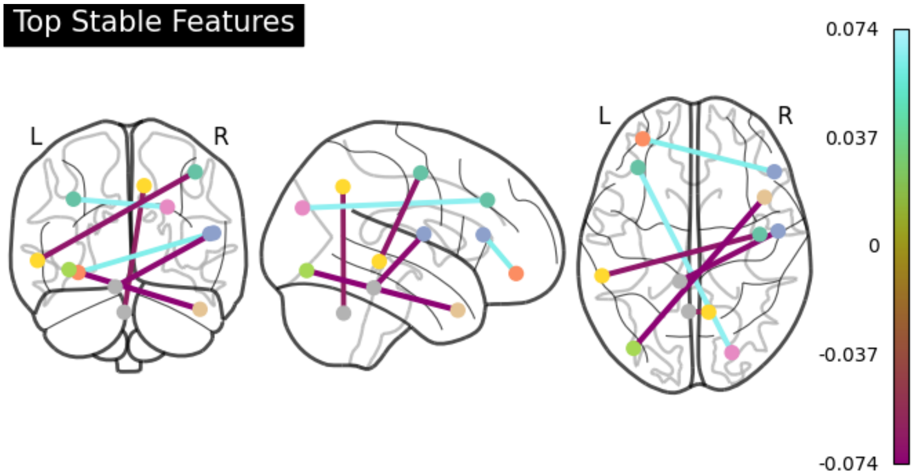


Figure 6.1: Connectome visualisation of the six most frequently occurring brain connectivity features across folds and models. Each edge represents a functional connection between two AAL regions that appeared repeatedly in the top 20 feature importance rankings (see Table 6.2).

7 Multi-Site Classification Performance on Graph Features

In the following, we present and analyse the performance of our classification models on graph features. As discussed in Section 3.2.2 we received a total of 29 datasets each containing the graph features for 871 subjects obtained using various inference and covariance estimation methods. We present and analyse the results on the graph features obtained with an adjacency inference method (`rSpecT)` with direct covariance estimation (dataset DF6). This method obtained the best performance balance between all metrics and scored the highest on average balanced accuracy across all models.

Unlike Pearson correlation, which captures pairwise temporal coactivation, `rSpecT` infers functional connectivity by modelling brain activity as the result of a diffusion process over a graph, capturing direct connections from the observable indirect relationships [19].

Performances on the other datasets can be seen in Appendix F.2.1 Table F.18–Table F.34.

7.1 Stratified Cross Validation

Table 7.1 presents the classification results on the graph-based features derived using the adjacency inference method, with direct covariance estimation. This dataframe contained only edge weights as features. When it comes to sensitivity, only DT and RF perform below the baseline, interestingly they are both tree-based models. Every classifier performs better than the baseline on all other metrics, by a great margin. Overall, the SVM model performs the best out of all models and outperforms the baseline by +11.2% for sensitivity, +5.8% for specificity, +12.8% for AUROC, +8.3% for accuracy, and +8.5% for balanced accuracy. RF achieved a very high specificity of 72.0%. However, the sensitivity of this model is very low (39.5%).

Table 7.1: Multisite classification performance on DF6.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
DUM	45.4% ± 3.0%	56.0% ± 5.7%	49.7% ± 2.4%	51.1 ± 4.0%	50.7 ± 3.9%
LR	55.1% ± 2.9%	60.9% ± 4.6%	60.8% ± 4.9%	58.2% ± 3.6%	58.0% ± 3.5%
SVM	**56.6% ± 6.4%**	61.8% ± 3.3%	**62.5% ± 4.8%**	**59.4% ± 4.6%**	**59.2% ± 4.7%**
DT	44.2% ± 12.7%	64.1% ± 10.5%	52.7% ± 2.9%	54.9% ± 1.5%	54.1% ± 1.9%
RF	43.7% ± 7.4%	**69.6% ± 3.9%**	61.2% ± 3.0%	57.6% ± 3.9%	56.7% ± 4.1%
MLP	54.8% ± 9.6%	62.2% ± 4.8%	61.5% ± 7.4%	58.8% ± 6.3%	58.5% ± 6.5%
LDA	53.4% ± 4.7%	59.8% ± 4.0%	60.1% ± 5.3%	56.8% ± 3.8%	56.6% ± 3.8%
KNN	48.4% ± 6.0%	64.3% ± 5.2%	58.6% ± 2.6%	56.9% ± 2.6%	56.3% ± 2.7%

7.1.1 Evaluation per site

Table 7.2 presents the average performance of the classifiers on the best performing site (YALE) and on the worst performing site (STANFORD) based on the mean balanced accuracy. The per-site performance for all individual sites can be found in Appendix F.2.2. Our LR model performs exceptionally well on the best-performing site, with a sensitivity of 81.7%, AUROC of 86.8%, and balanced accuracy of 79.2%. However the standard deviations are very high, though LR shows a more stable sensitivity performance than most other models. The performance on Stanford is sometimes almost twice as low as the performance on YALE. On YALE, all of our models perform better than the baseline across all metrics. Regarding achieved sensitivity, LR, SVM, and LDA perform exceptionally well on Yale.

On Stanford, no classifiers managed to match balanced accuracy or perform better than the baseline. Performance on LEUVEN_2 (Table F.37) was also remarkably low, with the exception of RF who obtained a balanced accuracy of 66.2%. Our SVM model performed really well on sites OHSU, PITT, and UCLA_2 with their balanced accuracies all above 70%.

Table 7.2: Comparison between the best performing site (**YALE**) and worst performing site (**STANFORD**).

	YALE					**STANFORD**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
DUM	46.7±7.5	32.0±29.5	45.2±27.1	44.6±10.4	39.3±17.3	41.7±35.5	**56.3±51.5**	**53.1±25.8**	**51.7±11.1**	**49.0±12.0**
LR	**81.7±20.7**	76.7±32.5	**86.8±19.5**	**74.2±19.5**	**79.2±17.7**	39.6±31.5	37.5±25.0	49.0±27.7	41.5±18.6	38.5±21.3
SVM	70.0±21.7	52.7±39.2	73.7±18.5	59.1±18.8	61.3±26.4	39.6±31.5	50.0±40.8	35.4±33.6	49.8±21.1	44.8±19.7
DT	40.0±38.4	**86.0±21.9**	78.8±11.3	57.6±18.1	63.0±11.4	47.9±44.3	31.2±37.5	50.5±12.5	47.5±14.5	39.6±12.5
RF	43.3±30.3	82.7±16.7	74.3±25.2	53.4±23.2	63.0±12.0	20.8±14.4	37.5±25.0	45.1±25.0	31.0±7.6	29.2±10.2
MLP	68.3±41.0	72.7±30.0	78.5±25.6	64.5±24.4	70.5±27.0	45.8±36.3	43.8±31.5	28.6±30.1	49.6±18.3	44.8±21.3
LDA	71.7±18.3	76.0±25.1	83.8±19.5	67.3±11.5	73.8±14.5	45.8±41.7	37.5±25.0	52.1±24.7	44.6±23.7	41.7±26.4
KNN	38.3±31.0	68.0±29.5	56.2±22.0	46.8±10.5	53.2±16.9	**60.4±42.7**	25.0±28.9	51.6±23.0	49.6±18.3	42.7±22.1

7.1.2 Evaluation per sex

Table 7.3 presents the performance of the seven classifiers trained on the graph features in DF6, evaluated separately for female and male subjects.
For the female subgroup, all classifiers underperformed in terms of sensitivity compared to the baseline. However, specificity was improved significantly across all models, with the smallest gain seen in the LDA model (+13.7%). AUROC performance was markedly better for all models, as the baseline AUROC was particularly low. In terms of accuracy, all models outperformed the baseline, although margins were narrow for some (e.g., +1.6% for LDA). For balanced accuracy, most classifiers showed only marginal improvements over the baseline, with MLP achieving the largest gain (+7.5%).
For the male data, all classifiers outperformed the dummy model on all metrics, although DT barely does so, with a sensitivity only 0.9% above baseline and an AUROC improvement of just 2.2%.

Table 7.3: The average performance across folds on DF6 graph-based features, evaluated per sex.

	Female					**Male**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
DUM	**58.4±10.2**	51.7±9.3	48.3±11.3	54.5±6.1	55.1±5.9	43.6±3.8	56.8±8.1	49.6±2.6	50.5±5.9	50.2%±5.7
LR	46.4±5.8	66.7±8.4	56.9±6.1	59.0±6.5	56.5±5.8	56.4±3.9	59.3±6.7	60.7±6.8	58.0±4.8	57.9±4.7
SVM	43.0±12.3	70.1±1.5	62.8±9.0	59.9±4.8	56.5±6.2	**58.4±8.2**	59.7±4.4	**62.2±6.1**	**59.1±6.1**	**59.1±6.1**
DT	41.9±16.9	75.1±10.5	55.2±8.4	62.7±7.2	58.5±8.3	44.5±12.5	61.6±10.6	51.8±3.4	53.4±1.7	53.0±1.9
RF	35.5±7.0	**80.0±9.7**	**66.2±1.4**	63.1±5.1	57.7±5.6	44.9±8.7	**67.4±3.4**	60.1±3.7	56.7±4.8	56.2±5.0
MLP	48.8±4.9	76.5±3.9	62.3±5.1	**66.0±1.8**	**62.6±1.6**	55.9±11.1	58.6±7.1	60.6±9.3	57.4±7.6	57.3±7.8
LDA	40.7±10.9	64.5±8.5	58.2±4.5	55.6±8.4	52.6±8.5	55.2±5.8	58.6±5.9	60.1±7.0	57.0±4.8	56.9±4.8
KNN	46.4±13.3	69.3±12.2	57.7±6.2	60.4±9.8	57.8±9.4	48.7±5.8	63.0±5.0	58.4±3.2	56.1±3.0	55.8±3.1

Across all classifiers, sensitivity was consistently higher for male subjects compared to female subjects. In contrast, specificity was consistently higher for female subjects compared to male subjects across all classifiers, with MLP, DT, and RF showing the largest gaps. For AUROC and balanced accuracy, the results varied per classifier, but female subjects often had a slight advantage. The MLP classifier achieved the highest balanced accuracy for female subjects (62.6%) and performed comparably on male subjects (57.3%). For both sexes, that is approximately 7% higher than the established baselines. The DT model had particularly low AUROC and balanced accuracy scores on male data, highlighting its weaker performance in that subgroup.

7.1.3 Evaluation per age

Table 7.4 highlights the best performing age group, 12–18 years, and the worst performing age group, 0–11 years based on their mean balanced accuracy. The complete results of all age groups can be found in Table F.45 andF.46.
The 12–18 group tends to have higher performance across all metrics. MLP, LR, and SVM show particularly strong average performance, with an AUROC score above 65% for all. LR achieved an accuracy of 62.2% and a balanced accuracy of 62.1%, while MLP reached the highest AUROC of 67.5%.
In contrast, the youngest age group showed notably lower performance. Most classifiers hovered around chance-level accuracy and balanced accuracy, with high standard deviations across folds (up to 15.8% for SVM sensitivity and 16.8% for LR AUROC). This instability may be attributed to the greater variability in early brain development, which may obscure patterns relevant to ASD classification.
Interestingly, the DT model performed relatively well in the 0–11 group, achieving the highest balanced accuracy (54.1%) and accuracy (56.2%) among the classifiers. This may indicate that simple, rule-based

classifiers are better suited for capturing the patterns present in early childhood, though overall performance remains limited. All classifiers surpassed the stratified dummy baseline (with a balanced accuracy of 46.7%), demonstrating limited but real predictive ability.
Across all metrics, the standard deviations in the 0–11 group are nearly double compared to the 12–18 group.

Table 7.4: The average performance across folds on graph features, evaluated per age group. Comparison between best-performing age group **12-18** (left) and worst-performing age group **0-11** (right).

	12-18					**0-11**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
DUM	47.1±5.0	56.2±6.6	49.3±7.1	51.8±5.3	51.6±5.3	40.2±12.8	53.1±12.3	49.3±7.3	47.5±10.9	46.7±11.2
LR	**58.7±3.0**	65.4±5.4	66.3±4.0	62.2±3.0	62.1±2.7	50.9±14.2	51.4±10.2	54.7±16.8	51.2±11.3	51.2±11.7
SVM	57.8±7.2	65.8±6.7	65.5±7.3	62.0±6.1	61.8±6.3	**57.2±15.8**	50.8±7.6	56.9±12.2	52.7±9.1	54.0±9.4
DT	43.3±12.9	64.3±11.0	52.0±4.6	54.0±3.3	53.8±3.6	40.6±14.7	**67.6±11.3**	**58.0±5.8**	**56.2±5.2**	**54.1±5.7**
RF	42.7±8.5	**71.0±3.5**	63.3±4.2	57.7±4.5	56.8±5.2	47.1±13.9	58.8±11.8	55.9±9.6	52.7±8.6	53.0±8.0
MLP	58.0±14.0	66.8±6.5	**67.5±9.5**	**62.8±8.9**	**62.4±9.5**	52.2±13.4	50.4±7.0	52.6±13.2	51.4±8.0	51.3±8.9
LDA	56.6±5.2	62.8±7.2	65.5±4.7	59.8±4.3	59.7±4.0	54.3±13.4	51.8±9.8	54.2±15.1	52.9±9.7	53.1±10.2
KNN	46.8±8.2	65.5±7.3	60.5±5.6	56.6±4.6	56.2±5.0	53.0±10.3	55.0±8.7	54.6±8.1	53.7±8.8	54.0±8.2

7.2 Cross-site Generalisation with LOGO CV

Table 7.5 shows the balanced accuracy obtained during LOGO CV. LR shows the best generalisation capability to unseen sites. LDA shows promising results, but struggles with LEUVEN_2 and MAXMUN. It appears that all models struggle with correctly classifying between ASD and TC for test site LEUVEN_2.
LR, MLP, LDA, and KNN achieved very high balanced accuracies on test site OHSU. Another test site that all models performed well on was UM_2. Almost all classifiers obtained a balanced accuracy of 70% or higher, with the exception of MLP and KNN who obtained balanced accuracies of 64.5% and 60.3%, respectively.
Of our classifiers, KNN struggled the most with generalising to unseen sites, obtaining a mean balanced accuracy 54.4% (±9.4%). However, on LEUVEN_2, on which all models struggled, KNN performed the best here. All in all, LR and LDA show best generalisation capability. SVM looks very promising too, but performs poorly on LEUVEN_1 and moderately on MAXMUN, PITT, and SBL. In addition to LEUVEN_2, the models also have difficulty with correctly identifying ASD in test site SBL.
Almost all models perform below 50%, with the exception of LDA and RF with balanced accuracies of 54.2% and 64.3%, respectively.

Table 7.5: Balanced Accuracy obtained by each classifier for each test site during LOGO CV on graph features. N indicates the number of ASD and TC samples in each site, which also corresponds to the size of the test set for that fold in LOGO CV, where each group (site) is left out once for testing.

	Balanced Accuracy [%]							**N**	
	LR	SVM	DT	RF	MLP	LDA	KNN	ASD	TC
CALTECH	65.0	**70.0**	60.0	35.0	55.0	60.0	35.0	5	10
CMU	**65.0**	55.0	48.3	58.3	46.7	56.7	38.3	6	5
KKI	56.5	52.4	**60.1**	52.4	58.3	58.3	50.0	12	21
LEUVEN_1	**64.3**	39.3	**64.3**	50.0	50.0	60.7	46.4	14	14
LEUVEN_2	42.7	58.3	50.0	53.1	44.8	45.8	**59.4**	12	16
MAXMUN	52.2	45.9	**64.6**	55.2	50.7	48.8	60.4	19	27
NYU	59.7	**65.5**	54.7	62.3	58.2	61.7	53.8	74	98
OHSU	72.1	56.1	59.9	52.9	**84.3**	71.8	72.1	12	13
OLIN	60.7	57.1	50.0	**64.3**	**64.3**	**64.3**	64.3	14	14
PITT	54.3	47.6	**61.7**	56.6	52.2	50.3	55.6	24	26
SBL	46.4	49.4	50.0	**64.3**	50.0	54.2	49.4	12	14
SDSU	**67.1**	53.0	53.6	52.0	55.6	60.9	64.5	8	19
STANFORD	56.7	**64.7**	56.7	55.4	64.4	60.9	56.1	12	13
TRINITY	50.4	**61.6**	48.1	57.6	42.4	52.4	45.1	19	25
UCLA_1	66.6	65.3	60.7	57.8	62.6	**69.8**	59.5	37	27
UCLA_2	56.8	61.8	52.7	**71.8**	61.4	56.8	65.9	11	10
UM_1	54.8	58.1	52.9	57.6	**61.3**	53.3	58.3	34	52
UM_2	75.1	**81.3**	72.3	70.3	64.5	71.2	60.3	13	21
USM	58.7	56.8	58.0	55.4	50.2	**60.1**	42.4	43	24
YALE	**75.5**	66.0	43.4	55.4	62.3	73.2	52.0	22	19
Mean±std	60.0±8.7	58.3±9.2	56.1±6.8	56.9±7.6	57.0±9.2	59.6±7.5	54.4±9.4		

7.3 Comparison with Pearson Correlation Features

Table 7.6 presents an overview of the balanced accuracies obtained by each feature extraction method in the multisite setting. In general, higher performance was obtained using Pearson correlation features. However, the training time on graph-based features was significantly faster than on Pearson correlation features due to the difference in dimensionality (190 features versus 6670 features).

Table 7.6: Balanced Accuracy obtained by each classifier for Pearson correlation and graph features during stratified cross-validation in the multisite setting.

	Balanced Accuracy [%]						
Method	LR	SVM	DT	RF	MLP	LDA	KNN
Pearson Correlation	62.4%	**64.4%**	57.0%	61.1%	64.3%	63.2%	58.9%
Graph features	58.0%	**59.2%**	54.1%	56.7%	58.5%	56.6%	56.3%

In addition to an overall assessment, results of both feature extraction methods were evaluated per site, per sex, and per age. Regarding the per site evaluation, both Pearson correlation and Graph-theoretical features struggled with correctly classifying ASD on Stanford. Interestingly, while Pearson correlation classification highly underperformed on OHSU, graph feature-based classification on OHSU showed superior performance with LR, SVM, and KNN with balanced accuracies of 65.0% for LR and KNN, and SVM achieving a balanced accuracy as high as 70.8%. Another site that was moderately difficult for both feature extraction methods, was classification on KKI, with balanced accuracies ranging from 51.0% to 64.7% for Pearson correlation and 45.2% to 69.8% for graph features. KNN with graph features performed quite good on KKI with an accuracy of 72.2% and balanced accuracy of 69.8% balanced accuracy. However, its sensitivity was on the low side being just 45.8%. LR and SVM score high balanced accuracies on PITT for both Pearson correlation (77.4% and 71.7%, respectively) and graph features (70.7% and 71.6%, respectively).

For per sex evaluation, MLP obtained highest balanced accuracy on the female subset for both Pearson correlation and graph-based features. In general, the Pearson correlation features performed better than the graph features for both male and female subsets.

For per age evaluation, both feature extraction methods obtained their best performance on the 12–18 age group and worst performance on the 0–11 age group. In general, the graph features performed worse than Pearson correlation features for both the 12–18 age group and the 0–11 age group. There is one performance metric where the graph features did manage to outperform the Pearson correlation features, namely for the sensitivity of the age group 0–11 (+4.8% for SVM, +4.2% for RF, and +0.3% for KNN).

Lastly, we compare the cross-site generalisation results of the Pearson correlation features and graph features. Table 7.7 shows a comparison of the mean balanced accuracy achieved by each classifier during LOGO CV using the two different feature extraction methods. Across all classifiers, the ability to generalise to unseen sites decreased when using graph features. However, the graph features method shows slightly more stable results. There were two instances where graph features outperformed Pearson correlation features, namely for test sites OHSU and UM_2 with sharp increases in balanced accuracies on test site UM_2 for all models, except MLP. For test site OHSU, balanced accuracy increased significantly in the LR, MLP, LDA, and KNN models. For test site Caltech, LR, SVM, DT, and LDA performed significantly better on graph features with increased balanced accuracy of +20%, +25%, +15%, and +20%, respectively. However, RF and MLP dropped significantly in performance on this test site with -20% and -10%, respectively. KNN showed relatively little difference (-5%).

Table 7.7: Balanced accuracy obtained by each classifier for Pearson correlation and graph features during LOGO CV.

	Balanced Accuracy (mean±std) [%]						
Method	LR	SVM	DT	RF	MLP	LDA	KNN
Pearson Correlation	61.3±11.9	61.8±11.7	57.5±10.7	60.4±10.8	**63.9±10.8**	59.7±8.9	56.2±11.3
Graph features	**60.0±8.7**	58.3±9.2	56.1±6.8	56.9±7.6	57.0±9.2	59.6±7.5	54.4±9.4

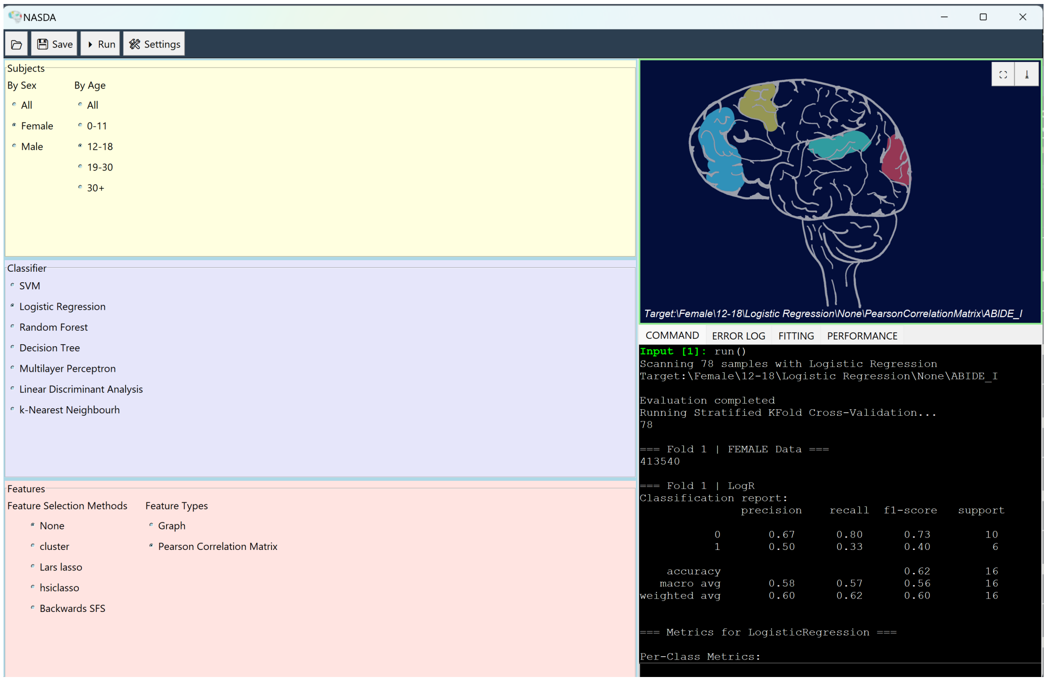
7.4 Discussion

12 of the 29 received graph-based datasets could not be used for testing and training. These data frames contained too many features (∼96% of total features) with 0 variance across subjects, thus having no discriminative power. These data frames were highlighted with an asterisk in Table C.1. In addition, graph-based datasets DF9–DF14 were inferred with the `rLogSpect` method, but without the proper tuning of the threshold and alpha parameter [19]. These data frames were highlighted with double asterisks in Table C.1. We obtained our best performance using dataset DF6, generated from an adjacency inference method with direct covariance estimation. Unfortunately, this feature dataframe only contained edge weight features. So, for now, the graph features do not outperform or match the performance obtained with Pearson correlation features, but we hope that the addition of other graph-based features to DF6, such as centrality features, clustering coefficients, and Laplacian eigenvalues, will improve the classification performance significantly, as the study in [11] reported an accuracy of 77.7% and 73.3% with their LDA and KNN model, respectively. With the current obtained graph-based dataframes, the classifiers have failed to satisfy requirements **NFR1-NFR6**. However, with this method, they do adhere to requirement **NFR7**.

Due to the delayed availability of the graph-derived features, a thorough feature importance analysis could not be performed within the current timeframe, marking it as a valuable focus for future research. The Feature Design subgroup transformed the resting-state fMRI time series from the 116 regions of the AAL atlas into 20 canonical resting-state networks (RSNs) defined by Smith *et al.* [46]. Since Pearson correlation and `rSpectT`-based FC rely on different assumptions about how connections between brain regions are established, comparing their feature importance may offer valuable insight into whether the two methods agree on important regions and connections, or whether `rSpecT` highlights additional connections as important in the decision-making of the classifiers.

8 NASDA (GUI)

The developed prototype, named NASDA (Neuroimaging Autism Spectrum Disorder Analyser), provides an interactive graphical user interface (GUI) designed to provide user-controlled setup and execution of machine learning pipelines on rs-fMRI data under a GNU Affero General Public License v3.0, shown in Figure 8.1.



Subjects

By Sex: ○ All ● Female ○ Male

By Age: ○ All ○ 0-11 ● 12-18 ○ 19-30 ○ 30+

Classifier: ○ SVM ● Logistic Regression ○ Random Forest ○ Decision Tree ○ Multilayer Perceptron ○ Linear Discriminant Analysis ○ k-Nearest Neighbourh

Features

Feature Selection Methods: ● None ○ cluster ○ Lars lasso ○ hsiclasso ○ Backwards SFS

Feature Types: ○ Graph ● Pearson Correlation Matrix

Target:\Female\12-18\Logistic Regression\None\PearsonCorrelationMatrix\ABIDE_I

```
Input [1]: run()
Scanning 78 samples with Logistic Regression
Target:\Female\12-18\Logistic Regression\None\ABIDE_I

Evaluation completed
Running Stratified KFold Cross-Validation...
78

=== Fold 1 | FEMALE Data ===
413540

=== Fold 1 | LogR
Classification report:
              precision    recall  f1-score   support

           0       0.67      0.80      0.73        10
           1       0.50      0.33      0.40         6

    accuracy                           0.62        16
   macro avg       0.58      0.57      0.56        16
weighted avg       0.60      0.62      0.60        16


=== Metrics for LogisticRegression ===

Per-Class Metrics:
```

Figure 8.1: Open-source graphical user interface prototype to control the Neuroimaging Autism Spectrum Disorder Analyser (NASDA) pipeline.

8.1 Purpose and Intended Audience

The NASDA prototype provides an accessible graphical user interface designed specifically for researchers and graduate students in neuro- and computational psychiatry. Its primary objective is to improve efficiency in exploratory ASD studies on rs-fMRI data by supporting subgroup filtering, feature selection, and model evaluation into a single, coherent workflow.
The interface allows users to interactively test hypotheses about demographic effects, regional brain patterns, and classifier performance. The integrated visualisation and command console promote transparency and traceability, supporting well-organised and rigorous research.

This tool is not developed as an over-the-counter diagnostic system but rather as a research aid: it enables domain experts to efficiently study the predictive value of graph-based and correlation-based brain features and supports the user to form new well-grounded hypotheses.
In summary, NASDA aims to bridge the gap between theoretical algorithms and the practical neuroimaging application, serving as an intuitive yet rigorous platform for reproducible, data-driven investigation.

8.2 System Functionalities

The prototype is structured into five primary functional modules: Subject Selection, Classifier Selection, Feature Configuration, Result Visualisation and Logging, and Performance Analysis and Output each described in detail below. All features have been programmed in such a way that anyone could adjust the program easily to their wishes for any future works. For a more in depth documentation, refer to Appendix G.

8.2.1 Subject Selection

The Subject Selection module enables dynamic subsetting of the dataset, by default ABIDE I, based on demographic attributes. Users can independently filter the available samples by sex (options: All, Female, Male) and age group (options: All, 0–11, 12–18, 19–30, 30+). Selection triggers an immediate update of the data partitions used for model training and evaluation. This functionality allows researchers to investigate demographic-specific patterns in functional brain connectivity and satisfies **FR3**.

8.2.2 Classifier Selection

The Classifier Selection panel provides a range of supervised learning algorithms commonly employed in neuroimaging classification tasks. Available options include: Support Vector Machine (SVM), Logistic Regression, Random Forest, Decision Tree, Multilayer Perceptron (MLP), Linear Discriminant Analysis (LDA), and k-Nearest Neighbour (KNN). The selected classifier is used during execution to fit models to the current demographic subset and feature representation. With this implemented, **FR5** is satisfied.

8.2.3 Feature Configuration

The Feature Configuration module offers two sub-components: Feature Selection Methods and Feature Types. For feature selection, users may choose from methods such as cluster, Lars lasso, hsiclasso, Backwards Sequential Feature Selection (SFS), or opt to use all available features (with None). In parallel, the Feature Types sub-component allows the user to switch between a graph-based representation, leveraging graph signal processing techniques, and a conventional Pearson Correlation Matrix. One can also change the analysed graph inference method, by adjusting the Graph Path in the settings window. This modular design supports comparative studies on how different feature extraction methods affect model performance and satisfies **FR4** and **FR6**.

8.2.4 Visualisation and Logging

A dedicated Brain Overview pane provides a visual summary of the current configuration overlaid on a visual representation of the brain, highlighting regions involved in the classification. Supplementary buttons allow the user to export the visualisation or expand it to full-screen view. Additionally, a tabbed console region logs command executions, input parameters, real-time run status, and evaluation results. Separate tabs for error reporting and performance summaries are included to enhance future interpretability and debugging possibilities. These functionalities deliver on **FR2** and **FR1**.

8.2.5 Performance Analysis and Output

The NASDA prototype provides integrated performance reporting for each classification run. Upon user request, the system executes the selected classifier on the currently filtered dataset, producing predictions and probabilistic estimates if supported by the model. The `run()` function serves as the main interface for model fitting, prediction, and evaluation and leans into **FR1**.

Performance metrics, including per-class precision, recall, F1-score, accuracy, macro averages, weighted avg, specificity, sensitivity, AUROC, and balanced accuracy are computed and displayed directly in the COMMAND tab within the GUI. This real-time feedback allows users to immediately interpret the results of a specific demographic or feature configuration without requiring external post-processing scripts. The console log clearly distinguishes between overall and fold-specific results and preserves a running history for comparison across multiple experiments. With this, the GUI satisfies **FR7**.

8.3 Design Choices

The NASDA prototype was developed with explicit design decisions to balance flexibility, interpretability, and ease of use for research-oriented rs-fMRI classification studies. The following key design choices guided the system architecture and user interaction flow:

1. **Modular User Interface**

   The graphical interface divides core operations into intuitive, context-specific panels: Subjects, Classifier, Features, and a dedicated Brain Overview canvas. This modular layout enables researchers to systematically configure subgroup filters (e.g., by sex or age), choose from a suite of standard classifiers, and select feature processing pipelines, all within a single workspace. Consistent radio button controls ensure clear and mutually exclusive parameter selection, reducing configuration conflicts.

2. **Inline Visual Feedback**

   An interactive brain illustration is persistently displayed to reinforce the domain-specific focus on neuroimaging. Overlay text dynamically updates to reflect the current target demographic, the selected classifier, and the feature configuration, providing live updated information on the configuration settings without requiring external documentation.

3. **Integrated Command Console**

   The prototype embeds a live `COMMAND` tab. This console captures executed commands, logs status messages, and prints fold-wise evaluation metrics. Researchers can inspect real-time feedback, track experiment history, and validate intermediate results without leaving the GUI environment. This makes a centralised and dynamically easy-to-use research workspace.

4. **Flexible Data Handling**

   The system supports on-the-fly filtering and reloading of the working dataset, driven by phenotypic metadata such as sex and age. This design accommodates subgroup analyses, which are critical in neurodevelopmental and clinical research contexts. For graph-based pipelines, the framework seamlessly switches between precomputed graph features and correlation matrices.

5. **Extensible Backend**

   All analytical operations, including model training, prediction, and performance evaluation, are modularised through dedicated functions (e.g., `performCA()`, `runCV()`). This integration promotes maintainability and accommodates integration with hyperparameter tuning or other new novel feature extraction methods without major refactoring of the GUI.

Overall, these design choices prioritise transparency and reproducibility, aligning with the prototype's role and **NFR8** and **NFR9**.

9 Discussion

In this chapter, we revisit the research questions outlined in Chapter 1 and discuss how our findings address each.

Which classifiers and feature types offer the best trade-off between accuracy and interpretability?

Logistic Regression, SVM, and LDA applied to the Pearson correlation features offered the best trade-off between accuracy and interpretability with accuracies of 62.7%, 64.7%, and 68.2%, and balanced accuracies of 62.4%, 64.4%, and 63.2%, respectively making them both conform to **NFR1** and **NFR5**. While MLP achieved similar performance with an accuracy of 64.5% and a balanced accuracy of 64.3%, it is not inherently interpretable and can only be interpreted by post-hoc explanations.

How can we identify brain regions that are both consistently important in the prediction across classifiers and experimental folds and how are these identified regions supported by established neuroscientific findings in ASD?

We addressed how we identify brain regions that are both consistently important in the prediction across classifiers and experimental folds in Section 3.6 and Section 6.1. Here, we will address how established neuroscientific findings in ASD research support our identified regions.
Several of the most frequently identified connections in our feature importance analysis align with established neuroscientific findings in ASD. The connection between the left inferior occipital gyrus (Occipital_Inf_L) and the right middle temporal pole (Temporal_Pole_Mid_R). The under-connectivity observed in the inferior occipital gyrus aligns with findings from Bai *et al.* and Long *et al.*, who reported similar under-connectivity of this brain region in a no-task state [47], [48]. This region is part of the occipital complex, which is primarily responsible for object recognition [49]. The inferior occipital gyrus has been found to be related to the visual function of processing faces, a function that is frequently disrupted in ASD [50]. Although prior studies have not specifically reported under-connectivity between the inferior occipital gyrus and the middle temporal pole in ASD, Olsen *et al.* discuss the role of the temporal pole in face processing [51].
The right precuneus, a key region of the default mode network involved in self-referential-processing and social processing, has been shown to exhibit abnormal connectivity in ASD [52]. Its recurrent pairing with vermis 9, part of the cerebellum, supports emerging evidence that the cerebellum may contribute to cognitive functions and the social brain [53], [54].
Although no direct link has been reported between the precentral gyrus and Heschl's gyrus, the broader primary auditory cortex, which includes Heschl's gyrus, is frequently linked to ASD. One study revealed that the right precentral gyrus (primary motor cortex) was less connected to the auditory cortices in ASD [55]. This is consistent with our model's negative feature weight for the connectivity between these regions, suggesting reduced functional connectivity in individuals with ASD compared to TC.
These findings suggest that the features identified as most important by our models are not arbitrary, but instead reflect neurologically meaningful altered patterns that have been observed in ASD research. By aligning feature importance results with established literature, this analysis strengthens the interpretability and credibility of the results and supports the potential of these connections as candidate biomarkers for further investigation in ASD.

How can we develop an intuitive interface to control the pipeline and visualise the resulting ROIs and model performance to support further neuroscientific analysis?

An open source GUI has been constructed to support an interactive and intuitive pipeline that can visualise and document model performances and the ROIs involved in the classification process. The GUI has been implemented to satisfy all functional requirements (**FR1-FR7**), **NFR8** and **NFR9**, as discussed in Chapter 8.

How can we improve the generalisability of our classifiers?

To improve the generalisability of our classifiers, we focussed primarily on two strategies: harmonisation on Pearson correlation features in the multi-site classification setting and the incorporation of graph-based features. First, it is important to distinguish between generalisability to unseen samples within known sites, which we assessed via stratified K-fold cross-validation, and generalisability to unseen sites. We evaluated the effect of harmonisation using NeuroHarmonize, a ComBat-based method designed to mitigate site-specific variability.

While harmonisation led to a modest improvement in balanced accuracy for our LR model and a rougly 2% increase in sensitivity for the MLP and KNN classifiers, these benefits were limited. Moreover, harmonisation introduced instability in model performance across folds. A notable limitation is that NeuroHarmonize requires all test sites to be represented in the training data, preventing its use for true unseen-site generalisability evaluation.

Secondly, we investigated whether graph-based features designed by the Feature Design subgroup could enhance generalisability by capturing more robust functional connectivity patterns. However, our current feature set was limited to only edge weights and did not include more complex graph metrics such as degree centrality, clustering coefficients, or Laplacian eigenvalues, which have been shown to be informative in prior studies [11]. Consequently, we could not fully assess the potential of graph-based features, but our initial results suggest that current edge-weight-only features do not improve generalisability and rather decrease the ability to generalise to unseen samples and unseen sites by a noticeable margin. We anticipate that incorporating more advanced graph features may offer improvements.

Additionally, given the high dimensionality of our feature space for Pearson correlation features (6670 features) relative to the number of subjects (871), feature selection could be vital to reducing overfitting and boosting performance.

To conclude, neither harmonisation nor graph features improved the generalisability of our classification models. However, when using traditional Pearson correlation features without harmonisation, our LR, SVM, and MLP models all met the non-functional requirements **NFR1-NFR7**. The LDA model also performed well, although its sensitivity standard deviation slightly exceeded the threshold specified in **NFR6**.

What evaluation methods best capture meaningful classifier performance beyond average accuracy, including site-specific, sex-based, or age-based assessments?

All per-subset evaluations offered valuable insights. Site-specific evaluation has shown us that performance disparities seen in some sites, like Stanford and OHSU, could be explained by differences in scan parameters and protocols.

Sex-specific evaluation revealed that different classifiers perform better depending on the subject's sex. Notably, our MLP model consistently achieved the highest performance across all metrics for female subjects in the multisite setting, reaching a balanced accuracy of 70.7% using Pearson correlation features, significantly outperforming the second-best model, which achieved 65.2%. Additionally, the MLP model showed a much higher sensitivity on female subjects (62.8% compared to 54.4% for the runner-up). In contrast, for male subjects, all models performed similarly, with no single model clearly outperforming the others.

Lastly, age-specific evaluation indicated that classifying ASD in the 0–11 years age group is substantially more challenging than in older age groups. This difficulty may stem from greater variability in connectivity patterns within this younger cohort, as they may exhibit both over- and under-connectivity, whereas adolescents (12–18) tend to show more consistent under-connectivity [44]. These findings suggest that functional connectivity features alone may have limited discriminative power for detecting ASD in young children, highlighting the need for future research to explore alternative or complementary biomarkers tailored to this age group.

Overall, subgroup-specific evaluation methods provide a richer, more detailed understanding of classifier performance than aggregate metrics alone, enabling targeted improvements and better interpretation in the context of ASD classification.

10 Conclusion and Future Work

10.1 Conclusion

This research focussed on the implementation and evaluation of a wide range of classifiers for the classification task of Autism Spectrum Disorder (ASD) detection. Seven classifiers were implemented and subjected to rigorous testing and evaluation on two types of input features: Pearson correlation features and graph-based features. Among these, Logistic Regression (LR), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) successfully satisfied all performance-related non-functional requirements (**NFR1-NFR7**) when evaluated on the Pearson correlation features. These three models obtained balanced accuracies of 62.4%, 64.4%, and 64.3%, respectively. Linear Discriminant Analysis (LDA) came close to meeting these requirements, falling short only on **NFR6** due to a standard deviation of 5.5%.

In contrast, the classifiers failed to meet the same criteria when evaluated on the graph-based features. This performance gap can largely be attributed to the limited informativeness of the graph features, which consisted solely of edge weights and lacked richer topological or contextual information. This outcome emphasises the critical importance of well-designed feature representations in machine learning pipelines, as model performance is inherently dependent on the quality of the input data.

This study is part of a larger interdisciplinary research aimed at improving the detection of ASD and discovering reliable neurological biomarkers. In support of this broader goal, we performed an in-depth feature importance analysis on the Pearson correlation features. This analysis identified several brain region connections that were consistently ranked as highly important across multiple folds and both LR and SVM classifiers. Notably, connections involving the inferior occipital gyrus, middle temporal pole, precuneus, and cerebellum appeared repeatedly, reflecting brain areas previously implicated in ASD-related research [56]-[55]. By linking these discriminative features to established neuroscientific findings, our results not only enhance model interpretability but also strengthen the case for these connections as potential biomarkers.

Lastly, we developed NASDA (Neuroimaging Autism Spectrum Disorder Analyser), an intuitive visual tool aimed to support neuroscientists in rs-fMRI analyses for research into atypical connectivity patterns in individuals with ASD. NASDA fulfilled all functional and non-functional requirements with Logistic Regression (the recommended model) for Pearson correlation features.

10.2 Future Work

While this study has demonstrated the effectiveness of several classifiers and provided insight into meaningful brain connectivity features for ASD classification, several promising directions remain for further exploration and improvement.

Firstly, the graph features pipeline can be significantly enriched by incorporating more advanced graph descriptions, such as Laplacian eigenvalues, nodal centralities, or community structure metrics, may better capture the topological nuances of brain connectivity.

Secondly, hyperparameter tuning for KNN and LDA may improve performance. In addition, to improve the computation speed during hyperparameter tuning, it might be worthwhile to parallelise this process. Moreover, experimenting with ensemble methods such as AdaBoost or histogram-based gradient boosting (HGBT) may provide performance gains and resilience to noise.

Another key aspect for future work is model interpretability and feature importance estimation. Applying SHAP (SHapley Additive exPlanations) to the MLP and extending feature importance analysis to the refined graph features, would provide deeper insights into which brain regions and connections drive classification decisions.

From a methodological perspective, exploring domain adaptation and multitask learning could make models more generalisable across sites, scanners, or demographic subgroups, addressing the known heterogeneity in ASD presentations. Here, parallel computing could again be utilised to handle the increased computational burden that comes with more complex models.

Lastly, future studies should consider incorporating factors such as handedness or ASD subtypes, as these may influence connectivity patterns and contribute to a more nuanced understanding of ASD-related neural biomarkers. Developing robust methods to address minority representation within these subgroups will be essential to ensure that findings are comprehensive, representative, and biologically meaningful.

Bibliography

[1] Statistics Netherlands (CBS), "3 percent of the population report having autism spectrum disorder," 2025, accessed: 9 June 2025. [Online]. Available: https://www.cbs.nl/en-gb/news/2025/14/3-percent-of-the-population-report-having-autism-spectrum-disorder

[2] M. J. Hollocks, J. W. Lerh, I. Magiati, R. Meiser-Stedman, and T. S. Brugha, "Anxiety and depression in adults with autism spectrum disorder: a systematic review and meta-analysis," *Psychol. Med.*, vol. 49, no. 4, pp. 559–572, Mar. 2019.

[3] american psychiatric, *Diagnostic and statistical manual of mental disorders, fifth edition, text revision (DSM-5-TR®)*, P. R. Muskin, A. L. Dickerman, A. Drysdale, and C. C. Holderness, Eds. Arlington, TX: American Psychiatric Association Publishing, Apr. 2022.

[4] S.-J. Hong, R. Vos de Wael, R. A. I. Bethlehem, S. Lariviere, C. Paquola, S. L. Valk, M. P. Milham, A. Di Martino, D. S. Margulies, J. Smallwood, and B. C. Bernhardt, "Atypical functional connectome hierarchy in autism," *Nat. Commun.*, vol. 10, no. 1, p. 1022, Mar. 2019.

[5] 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," *Mol. Psychiatry*, vol. 19, no. 6, pp. 659–667, Jun. 2014.

[6] 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, 2013.

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

[8] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the abide dataset," *NeuroImage: Clinical*, vol. 17, p. 16–23, 2018.

[9] X. Yang, M. S. Islam, and A. M. A. Khaled, "Functional connectivity magnetic resonance imaging classification of autism spectrum disorder using the multisite ABIDE dataset," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, May 2019.

[10] T. Eslami, V. Mirjalili, A. Fong, A. R. Laird, and F. Saeed, "Asd-diagnet: A hybrid learning approach for detection of autism spectrum disorder using fmri data," *Frontiers in Neuroinformatics*, vol. 13, Nov. 2019.

[11] S. Mostafa, L. Tang, and F.-X. Wu, "Diagnosis of autism spectrum disorder based on eigenvalues of brain networks," *IEEE Access*, vol. 7, pp. 128 474–128 486, 2019.

[12] 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, p. 27–37, Feb. 2020.

[13] M. Tang, P. Kumar, H. Chen, and A. Shrivastava, "Deep multimodal learning for the diagnosis of autism spectrum disorder," *Journal of Imaging*, vol. 6, no. 6, p. 47, Jun. 2020.

[14] A. Kazeminejad and R. C. Sotero, "The importance of anti-correlations in graph theory based classification of autism spectrum disorder," *Frontiers in Neuroscience*, vol. 14, Aug. 2020.

[15] M. A. Reiter, A. Jahedi, A. R. Jac 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 Comput. Appl.*, vol. 33, no. 8, pp. 3299–3310, Apr. 2021.

[16] 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*, vol. 15, no. 1, pp. 273–289, Jan. 2002.

[17] R. C. Craddock, G. A. James, P. E. Holtzheimer, 3rd, X. P. Hu, and H. S. Mayberg, "A whole brain fMRI atlas generated via spatially constrained spectral clustering," *Hum. Brain Mapp.*, vol. 33, no. 8, pp. 1914–1928, Aug. 2012.

[18] C. Craddock, S. Sikka, B. Cheung, R. Khanuja, S. S. Ghosh, C. Yan, Q. Li, D. Lurie, J. Vogelstein, R. Burns, S. Colcombe, M. Mennes, C. Kelly, A. Di Martino, F. X. Castellanos, and M. Milham, "Towards automated analysis of connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC)," *Frontiers in Neuroinformatics*, no. 42, 2013.

[19] J. Groenenberg and G. Myburg, "Graph-based feature extraction for functional connectivity analysis in rs-fMRI of individuals with ASD and controls," 2025, BSc thesis, unpublished, Department of Microelectronics, TU Delft.

[20] D. Collins, A. Zijdenbos, V. Kollokian, J. Sled, N. Kabani, C. Holmes, and A. Evans, "Design and construction of a realistic digital brain phantom," *IEEE Transactions on Medical Imaging*, vol. 17, no. 3, pp. 463–468, 1998.

[21] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI," *Magn. Reson. Med.*, vol. 34, no. 4, pp. 537–541, Oct. 1995.

[22] N. contributors, *nilearn*. [Online]. Available: https://github.com/nilearn/nilearn

[23] A. Yamashita, N. Yahata, T. Itahashi, G. Lisi, T. Yamada, N. Ichikawa, M. Takamura, Y. Yoshihara, A. Kunimatsu, N. Okada, H. Yamagata, K. Matsuo, R. Hashimoto, G. Okada, Y. Sakai, J. Morimoto, J. Narumoto, Y. Shimada, K. Kasai, N. Kato, H. Takahashi, Y. Okamoto, S. C. Tanaka, M. Kawato, O. Yamashita, and H. Imamizu, "Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias," *PLOS Biology*, vol. 17, no. 4, p. e3000042, Apr. 2019.

[24] R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I. M. Nasrallah, T. D. Satterthwaite, Y. Fan, L. J. Launer, C. L. Masters, P. Maruff, C. Zhuo, H. Völzke, S. C. Johnson, J. Fripp, N. Koutsouleris, D. H. Wolf, R. Gur, R. Gur, J. Morris, M. S. Albert, H. J. Grabe, S. M. Resnick, R. N. Bryan, D. A. Wolk, R. T. Shinohara, H. Shou, and C. Davatzikos, "Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan," *NeuroImage*, vol. 208, p. 116450, 2020.

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

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

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

[28] M. Yu, K. A. Linn, P. A. Cook, M. L. Phillips, M. McInnis, M. Fava, M. H. Trivedi, M. M. Weissman, R. T. Shinohara, and Y. I. Sheline, "Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data," *Human Brain Mapping*, vol. 39, no. 11, p. 4213–4227, Jul. 2018.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] D. C. Liu and J. Nocedal, “On the limited memory BFGS method for large scale optimization,” *Math. Program.*, vol. 45, no. 1-3, pp. 503–528, Aug. 1989.

[31] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[32] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning*, 2nd ed., ser. Springer series in statistics. New York, NY: Springer, Dec. 2009, pp. 436–437.

[33] ——, *The elements of statistical learning*, 2nd ed., ser. Springer series in statistics. New York, NY: Springer, Dec. 2009, pp. 109–110.

[34] X. Guo, H. Chen, Z. Long, X. Duan, Y. Zhang, and H. Chen, “Atypical developmental trajectory of local spontaneous brain activity in autism spectrum disorder,” *Sci. Rep.*, vol. 7, no. 1, Jan. 2017.

[35] M. Cao, J.-H. Wang, Z.-J. Dai, X.-Y. Cao, L.-L. Jiang, F.-M. Fan, X.-W. Song, M.-R. Xia, N. Shu, Q. Dong, M. P. Milham, F. X. Castellanos, X.-N. Zuo, and Y. He, “Topological organization of the human brain functional connectome across the lifespan,” *Dev. Cogn. Neurosci.*, vol. 7, pp. 76–93, Jan. 2014.

[36] Python Software Foundation, *Tkinter Standard GUI Toolkit*, Python Software Foundation, 2025, https://docs.python.org/3/library/tkinter.html.

[37] A. Clark and Contributors, “Pillow (pil fork) documentation,” https://pillow.readthedocs.io/, 2025, version 11.2.1.

[38] H. Haghighat, “A sex-dependent functional-effective connectivity model for diagnostic classification of autism spectrum disorder using resting-state fmri,” *Biomedical Signal Processing and Control*, vol. 85, p. 104837, 2023.

[39] J. Y. Namgung, J. Mun, Y. Park, J. Kim, and B. yong Park, “Sex differences in autism spectrum disorder using class imbalance adjusted functional connectivity,” *NeuroImage*, vol. 304, p. 120956, 2024.

[40] W. Wang and D. Sun, “The improved adaboost algorithms for imbalanced data classification,” *Information Sciences*, vol. 563, pp. 358–374, 2021.

[41] S. Mueller, D. Keeser, A. C. Samson, V. Kirsch, J. Blautzik, M. Grothe, O. Erat, M. Hegenloh, U. Coates, M. F. Reiser, K. Hennig-Fast, and T. Meindl, “Convergent findings of altered functional and structural brain connectivity in individuals with high functioning autism: A multimodal MRI study,” *PLoS One*, vol. 8, no. 6, p. e67329, Jun. 2013.

[42] T. D. Satterthwaite, D. H. Wolf, J. Loughead, K. Ruparel, M. A. Elliott, H. Hakonarson, R. C. Gur, and R. E. Gur, “Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth,” *Neuroimage*, vol. 60, no. 1, pp. 623–632, Mar. 2012.

[43] J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar, and S. E. Petersen, “Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion,” *Neuroimage*, vol. 59, no. 3, pp. 2142–2154, Feb. 2012.

[44] J. V. Hull, L. B. Dokovna, Z. J. Jacokes, C. M. Torgerson, A. Irimia, and J. D. Van Horn, “Resting-state functional connectivity in autism spectrum disorders: A review,” *Front. Psychiatry*, vol. 7, p. 205, 2016.

[45] C. Shi, X. Xin, and J. Zhang, “Domain adaptation using a Three-Way decision improves the identification of autism patients from multisite fMRI data,” *Brain Sci.*, vol. 11, no. 5, p. 603, May 2021.

[46] S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, and C. F. Beckmann, “Correspondence of the brain’s functional architecture during activation and rest,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 31, pp. 13 040–13 045, Aug. 2009.

[47] C. Bai, Y. Wang, Y. Zhang, X. Wang, Z. Chen, W. Yu, H. Zhang, X. Li, K. Zhu, Y. Wang, and T. Zhang, “Abnormal gray matter volume and functional connectivity patterns in social cognition-related brain regions of young children with autism spectrum disorder,” *Autism Res.*, vol. 16, no. 6, pp. 1124–1137, Jun. 2023.

[48] Z. Long, X. Duan, D. Mantini, and H. Chen, "Alteration of functional connectivity in autism spectrum disorder: effect of age and anatomical distance," *Sci. Rep.*, vol. 6, no. 1, May 2016.

[49] K. Grill-Spector, Z. Kourtzi, and N. Kanwisher, "The lateral occipital complex and its role in object recognition," *Vision Res.*, vol. 41, no. 10-11, pp. 1409–1422, 2001.

[50] W. Sato, T. Kochiyama, S. Uono, K. Matsuda, K. Usui, N. Usui, Y. Inoue, and M. Toichi, "Bidirectional electric communication between the inferior occipital gyrus and the amygdala during face processing," *Hum. Brain Mapp.*, vol. 38, no. 9, pp. 4511–4524, Sep. 2017.

[51] I. R. Olson, A. Plotzker, and Y. Ezzyat, "The enigmatic temporal pole: a review of findings on social and emotional processing," *Brain*, vol. 130, no. Pt 7, pp. 1718–1731, Jul. 2007.

[52] A. Padmanabhan, C. J. Lynch, M. Schaer, and V. Menon, "The default mode network in autism," *Biol. Psychiatry Cogn. Neurosci. Neuroimaging*, vol. 2, no. 6, pp. 476–486, Sep. 2017.

[53] S. S.-H. Wang, A. D. Kloth, and A. Badura, "The cerebellum, sensitive periods, and autism," *Neuron*, vol. 83, no. 3, pp. 518–532, Aug. 2014.

[54] L. Mapelli, T. Soda, E. D'Angelo, and F. Prestori, "The cerebellar involvement in autism spectrum disorders: From the social brain to mouse models," *Int. J. Mol. Sci.*, vol. 23, no. 7, p. 3894, Mar. 2022.

[55] K. C. Wilson, M. Kornisch, and T. Ikuta, "Disrupted functional connectivity of the primary auditory cortex in autism," *Psychiatry Res. Neuroimaging*, vol. 324, no. 111490, p. 111490, Aug. 2022.

[56] C. L. Keown, P. Shih, A. Nair, N. Peterson, M. E. Mulvey, and R.-A. Müller, "Local functional overconnectivity in posterior brain regions is associated with symptom severity in autism spectrum disorders," *Cell Rep.*, vol. 5, no. 3, pp. 567–572, Nov. 2013.

[57] R. W. Cox, "Afni: Software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical Research*, vol. 29, no. 3, pp. 162–173, 1996.

[58] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical Image Analysis*, vol. 5, no. 2, pp. 143–156, 2001.

[59] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, pp. 825–841, 2002.

[60] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[61] J. M. Tyszka, D. P. Kennedy, L. K. Paul, and R. Adolphs, "Largely typical patterns of resting-state functional connectivity in high-functioning adults with autism," *Cereb. Cortex*, vol. 24, no. 7, pp. 1894–1905, Jul. 2014.

[62] M. B. Nebel, S. E. Joel, J. Muschelli, A. D. Barber, B. S. Caffo, J. J. Pekar, and S. H. Mostofsky, "Disruption of functional organization within the primary motor cortex in children with autism," *Hum. Brain Mapp.*, vol. 35, no. 2, pp. 567–580, Feb. 2014.

A ABIDE I Dataset Details

Table A.1 and A.2 contain the subject information and scanner information of the ABIDE I dataset.

A.1 Preprocessing Pipeline (CPAC)

The ABIDE data was preprocessed using the CPAC pipeline. The steps included:

1. Slice time correction using AFNI's *3dTshift* [57].
2. Motion correct to the average image using AFNI's *3dvolreg* (two iterations).
3. Skull-strip using AFNI's *3dAutomask*.
4. Global mean intensity normalization to 10,000.
5. Nuisance signal regression was applied including:
    - motion parameters: 6 head motion parameters, 6 head motion parameters one time point before, and the 12 corresponding squared items
    - top 5 principal components from the signal in the white-matter and cerebro-spinal fluid derived from the prior tissue segmentations transformed from anatomical to functional space
    - linear and quadratic trends
6. Band-pass filtering (0.01-0.1Hz)
7. Functional images were registered to anatomical space with a linear transformation and then a white-matter boundary based transformation using FSL's *FLIRT* [58, 59] and the prior white-matter tissue segmentation from *FAST* [60].
8. The previous anatomical to standard space registration was applied to the functional data in order to transform them to standard space.

A.2 Subject Exclusion

The original ABIDE I dataset consisted of 1112 subjects. After preprocessing with CPAC, 871 subjects remained. A total of 241 subjects were excluded for the following reasons:

- **Information extraction error**: subjects were skipped due to an error occurring during the information extraction from the phenotypic file.
- **Missing FILE_ID entries**: subjects were skipped due to missing file names in the phenotypic file.
- **Framewise Displacement (FD) threshold**: subjects were excluded because they exceeded the FD threshold of 0.2.
- **Quality Check**: subjects that did not pass the quality assessment were excluded.

The final dataset contained 871 subjects (403 ASD, 468 controls).

Table A.1: Subject information of the raw ABIDE I dataset per acquisition site

Site	Abbreviation	N (Total)	N (ASD)	N (Control)	% Female	Mean Age ± SD (years)	Age range (years)
California Institute of Technology [61]	CALTECH	38	19	19	21.05	28.16±10.64	17.0-56.2
Carnegie Mellon University	CMU	27	14	13	22.22	26.59±5.69	19-40
Kennedy Krieger Institute [62]	KKI	55	22	33	23.64	10.1±1.33	8.0-12.8
Ludwig Maximilians University Munich [41]	MAXMUN	57	24	33	12.28	26.16±12.08	7-58
New York University Langone Medical Center	NYU	184	79	105	20.11	15.25±6.58	6.5-39.1
Olin, Institute of Living at Hartford Hospital	OLIN	36	20	16	13.89	16.81±3.49	10-24
Oregon Health and Science University	OHSU	28	13	15	0.0	10.81±1.87	8.0-15.2
San Diego State University	SDSU	36	14	22	19.44	14.41±1.84	8.7-17.2
Social Brain Lab	SBL	30	15	15	0.0	34.37±8.6	20-64
Stanford University	STANFORD	40	20	20	20.0	9.96±1.58	7.5-12.9
Trinity Centre for Health Sciences	TRINITY	49	24	25	0.0	17.18±3.64	12.0-25.9
University of California, Los Angeles 1	UCLA_1	82	49	33	13.41	13.16±2.31	8.4-17.9
University of California, Los Angeles 2	UCLA_2	27	13	14	7.41	12.46±1.5	9.8-16.5
University of Leuven 1	LEUVEN_1	29	14	15	0.0	22.59±3.55	18-32
University of Leuven 2	LEUVEN_2	35	15	20	22.86	14.16±1.42	12.1-16.9
University of Michigan 1	UM1	110	55	55	23.64	13.4±2.89	8.2-19.2
University of Michigan 2	UM2	35	13	22	5.71	15.96±3.32	12.8-28.8
University of Pittsburgh School of Medicine	PITT	57	30	27	14.04	18.9±6.88	9.5-35.2
University of Utah School of Medicine	USM	101	58	43	0.0	22.1±7.68	8.8-50.2
Yale Child Study Center	YALE	56	28	28	28.57	12.71±2.88	7.0-17.8

Table A.2: Scanner information of the acquisition sites

Site	Scanner Model	Field Strength	Slice Thickness (mm)	TR (ms)	TE (ms)	Voxel size (mm)	Measurements
CALTECH	SIEMENS MAGNETOM TrioTim syngo MR B17	3T	3.5	2000	30	3.5×3.5×3.5	150
CMU	SIEMENS MAGNETOM Verio syngo MR B17	3T	3.0	2000	30	3.0×3.0×3.0	240
KKI	Philips 3T Achieva	3T	3.0	2500	30	3.05×3.15×3.0	156
MAXMUN	SIEMENS MAGNETOM Verio syngo MR B17	3T	4.0	3000	30	3.0×3.0×4.0	120
NYU	SIEMENS MAGNETOM Allegra syngo MR 2004A	3T	4.0	2000	15	3.0×3.0×4.0	180
OLIN	SIEMENS MAGNETOM Allegra syngo MR 2004A	3T	4.0	1500	27	3.4×3.4×4.0	210
OHSU	SIEMENS MAGNETOM TrioTim syngo MR B17	3T	3.8	2500	30	3.8×3.8×3.8	82
SDSU	GE 3T MR750	3T	3.4	2000	30	3.4×3.4×3.4	180
SBL	Philips Intera 3T	3T	2.72	2200	30	2.75×2.75×2.72	-
STANFORD	GE SIGNA 3T	3T	4.5	2000	30	3.125×3.125×4.5	180
TRINITY	Philips 3T Achieva	3T	3.5	2000	28	3.0×3.0×3.5	150
UCLA	SIEMENS MAGNETOM TrioTim syngo MR B15	3T	4.0	3000	28	3.0×3.0×4.0	120
LEUVEN_1	Philips Intera 3T	3T	4.0	1667	33	3.59×3.59×4.0	250
LEUVEN_2	Philips Intera 3T	3T	4.0	1667	33	3.59×3.59×4.0	250
UM	3T GE Signa	3T	3.0	2000	30	3.438×3.438×3.0	300
PITT	SIEMENS MAGNETOM Allegra syngo MR A30	-	4.0	1500	25	3.1×3.1×4.0	200
USM	SIEMENS MAGNETOM TrioTim syngo MR B17	3T	3.0	2000	28	3.4×3.4×3.0	240
YALE	SIEMENS MAGNETOM TrioTim syngo MR B17	3T	4.0	2000	25	3.4×3.4×4.0	200

Table A.3: Subject information of the preprocessed ABIDE I dataset per acquisition site.

	N					
Site	Total	ASD	TC	% Female	Age (Mean±SD)	Age range
CALTECH	15	5	10	33.33	26.79±10.77	17.0–56.2
CMU	11	6	5	36.36	26.82±4.81	19.0–33.0
KKI	33	12	21	27.27	10.31±1.28	8.2–12.8
MAXMUN	46	19	27	8.70	26.5±10.63	7.0–58.0
NYU	172	74	98	20.93	15.33±6.6	6.47–39.1
OLIN	28	14	14	17.86	17.04±3.43	10–24
OHSU	25	12	13	0.0	10.81±1.75	8.0–15.23
SDSU	27	8	19	22.22	14.36±1.93	8.67–17.15
SBL	26	12	14	0.0	33.77±6.6	20.0–49.0
STANFORD	25	12	13	28.0	9.99±1.63	7.53–12.94
TRINITY	44	19	25	0.0	17.03±3.5	12.0–25.66
UCLA_1	64	37	27	14.06	13.35±2.38	8.36–17.94
UCLA_2	21	11	10	9.52	12.47±1.67	9.79–16.47
LEUVEN_1	28	14	14	0.0	22.43±3.51	18.0–32.0
LEUVEN_2	28	12	16	25.0	14.17±1.48	12.1–16.9
UM_1	86	34	52	29.07	13.77±2.96	8.2–19.2
UM_2	34	13	21	5.88	16.02±3.36	12.8–28.8
PITT	50	24	26	14.0	18.5±6.76	9.33–35.2
USM	67	43	24	0.0	22.59±8.36	8.77–50.22
YALE	41	22	19	39.02	13.31±2.64	7.0–17.75
TOTAL	871	403	468	16.53	16.94±7.58	6.47–58.0

B AAL Index to Region Mapping

- 2001 - Precentral_L,
- 2002 - Precentral_R,
- 2101 - Frontal_Sup_L,
- 2102 - Frontal_Sup_R,
- 2111 - Frontal_Sup_Orb_L,
- 2112 - Frontal_Sup_Orb_R,
- 2201 - Frontal_Mid_L,
- 2202 - Frontal_Mid_R,
- 2211 - Frontal_Mid_Orb_L,
- 2212 - Frontal_Mid_Orb_R,
- 2301 - Frontal_Inf_Oper_L,
- 2302 - Frontal_Inf_Oper_R,
- 2311 - Frontal_Inf_Tri_L,
- 2312 - Frontal_Inf_Tri_R,
- 2321 - Frontal_Inf_Orb_L,
- 2322 - Frontal_Inf_Orb_R,
- 2331 - Rolandic_Oper_L,
- 2332 - Rolandic_Oper_R,
- 2401 - Supp_Motor_Area_L,
- 2402 - Supp_Motor_Area_R,
- 2501 - Olfactory_L,
- 2502 - Olfactory_R,
- 2601 - Frontal_Sup_Medial_L,
- 2602 - Frontal_Sup_Medial_R,
- 2611 - Frontal_Med_Orb_L,
- 2612 - Frontal_Med_Orb_R,
- 2701 - Rectus_L,
- 2702 - Rectus_R,
- 3001 - Insula_L,
- 3002 - Insula_R,
- 4001 - Cingulum_Ant_L,
- 4002 - Cingulum_Ant_R,
- 4011 - Cingulum_Mid_L,
- 4012 - Cingulum_Mid_R,
- 4021 - Cingulum_Post_L,
- 4022 - Cingulum_Post_R,
- 4101 - Hippocampus_L,
- 4102 - Hippocampus_R,
- 4111 - ParaHippocampal_L,
- 4112 - ParaHippocampal_R,
- 4201 - Amygdala_L,
- 4202 - Amygdala_R,
- 5001 - Calcarine_L,
- 5002 - Calcarine_R,
- 5011 - Cuneus_L,
- 5012 - Cuneus_R,
- 5021 - Lingual_L,
- 5022 - Lingual_R,
- 5101 - Occipital_Sup_L,
- 5102 - Occipital_Sup_R,
- 5201 - Occipital_Mid_L,
- 5202 - Occipital_Mid_R,
- 5301 - Occipital_Inf_L,
- 5302 - Occipital_Inf_R,
- 5401 - Fusiform_L,
- 5402 - Fusiform_R,
- 6001 - Postcentral_L,
- 6002 - Postcentral_R,
- 6101 - Parietal_Sup_L,
- 6102 - Parietal_Sup_R,
- 6201 - Parietal_Inf_L,
- 6202 - Parietal_Inf_R,
- 6211 - SupraMarginal_L,
- 6212 - SupraMarginal_R,
- 6221 - Angular_L,
- 6222 - Angular_R,
- 6301 - Precuneus_L,
- 6302 - Precuneus_R,
- 6401 - Paracentral_Lobule_L,
- 6402 - Paracentral_Lobule_R,
- 7001 - Caudate_L,
- 7002 - Caudate_R,
- 7011 - Putamen_L,
- 7012 - Putamen_R,
- 7021 - Pallidum_L,
- 7022 - Pallidum_R,
- 7101 - Thalamus_L,
- 7102 - Thalamus_R,
- 8101 - Heschl_L,
- 8102 - Heschl_R,
- 8111 - Temporal_Sup_L,
- 8112 - Temporal_Sup_R,
- 8121 - Temporal_Pole_Sup_L,
- 8122 - Temporal_Pole_Sup_R,
- 8201 - Temporal_Mid_L,
- 8202 - Temporal_Mid_R,
- 8211 - Temporal_Pole_Mid_L,
- 8212 - Temporal_Pole_Mid_R,
- 8301 - Temporal_Inf_L,
- 8302 - Temporal_Inf_R,
- 9001 - Cerebelum_Crus1_L,
- 9002 - Cerebelum_Crus1_R,
- 9011 - Cerebelum_Crus2_L,
- 9012 - Cerebelum_Crus2_R,
- 9021 - Cerebelum_3_L,

- 9022 - Cerebelum_3_R,
- 9031 - Cerebelum_4_5_L,
- 9032 - Cerebelum_4_5_R,
- 9041 - Cerebelum_6_L,
- 9042 - Cerebelum_6_R,
- 9051 - Cerebelum_7b_L,
- 9052 - Cerebelum_7b_R,
- 9061 - Cerebelum_8_L,
- 9062 - Cerebelum_8_R,
- 9071 - Cerebelum_9_L,
- 9072 - Cerebelum_9_R,
- 9081 - Cerebelum_10_L,
- 9082 - Cerebelum_10_R,
- 9100 - Vermis_1_2,
- 9110 - Vermis_3,
- 9120 - Vermis_4_5,
- 9130 - Vermis_6,
- 9140 - Vermis_7,
- 9150 - Vermis_8,
- 9160 - Vermis_9,
- 9170 - Vermis_10

C List of Graph Features

C.1 Complete List of Graph Features

Node Specific Features

- Degree Centrality
- Eigenvector Centrality
- Clustering Coefficient
- Betweenness Centrality
- Closeness Centrality

Node to Node

- Edge weights

Global Features

- Average Clustering
- Diameter
- Graph energy
- Spectral entropy
- Mean Laplacian Eigenvalue
- Max Laplacian Eigenvalue
- Frobenius Norm (Laplacian Spectrum)
- Algebraic Connectivity
- Characteristic Path Length
- Smallworldness
- Modularity

C.2 Features Included in Each Dataframe

Table C.1: An overview of the features included in each dataframe received from the feature design subgroup [19]. DF1-DF8 features were derived from 20 ICA components with Smith ICA, whereas DF9-DF29 features were derived from 30 ICA components with group ICA.

ID	Inference	Cov. Estimation	#ICA	alpha	threshold	Features	# Features
DF1	Normalized Laplacian	direct	20	0.45	0.15	edge weights	190
DF2	Normalized Laplacian	VAR(k)	20	0.45	0.15	edge weights	190
DF3	Normalized Laplacian	Ledoit-Wolf shrinkage	20	0.45	0.15	edge weights	190
DF4	Normalized Laplacian	Glasso	20	0.45	0.15	edge weights	190
DF5	Normalized Laplacian	direct	20	0.0001	0.25	edge weights	190
DF6	Adjacency	direct	20	0.0001	0.10	edge weights	190
DF7	`rLogSpecT`	VAR(k)	20	15	0.2	edge weights	190
DF8	`rLogSpecT`	direct	20	15	0.2	edge weights	190
DF9**	`rLogSpecT`	sliding window	30	N/S	0.05	All, except edge weights	130
DF10**	`rLogSpecT`	VAR(k)	30	N/S	0.05	All, except edge weights	130
DF11**	`rLogSpecT`	direct	30	N/S	0.05	All, except edge weights	130
DF12**	`rLogSpecT`	Glasso	30	N/S	0.05	All, except edge weights	130
DF13**	`rLogSpecT`	Ledoit-Wolf shrinkage	30	N/S	0.05	All, except edge weights	130
DF14**	`rLogSpecT`	NVAR	30	N/S	0.05	All, except edge weights	130
DF15	Sample covariance	NVAR	30	N/A	0.05	All, except edge weights	130
DF16*	Sample covariance	direct	30	N/A	0.05	All, except edge weights	130
DF17*	Sample covariance	Glasso	30	N/A	0.05	All, except edge weights	130
DF18*	Sample covariance	Ledoit-Wolf shrinkage	30	N/A	0.05	All, except edge weights	130
DF19*	Sample covariance	VAR(k)	30	N/A	0.05	All, except edge weights	130
DF20*	Sample covariance	sliding window	30	N/A	0.05	All, except edge weights	130
DF21*	Partial correlation	direct	30	N/A	0.05	All, except edge weights	130
DF22	Partial correlation	NVAR	30	N/A	0.05	All, except edge weights	130
DF23*	Partial correlation	Glasso	30	N/A	0.05	All, except edge weights	130
DF24*	Partial correlation	Ledoit-Wolf shrinkage	30	N/A	0.05	All, except edge weights	130
DF25*	Partial correlation	VAR(k)	30	N/A	0.05	All, except edge weights	130
DF26*	Partial correlation	sliding window	30	N/A	0.05	All, except edge weights	130
DF27	Mutual information	N/A	30	N/A	0.05	All graph features	565†
DF28*	Pearson correlation	N/A	30	N/A	0.05	All graph features	565†
DF29*	Granger causality	N/A	30	N/A	0.05	All graph features	565†

* These dataframes contained too many zero variance features ($\sim$ 96%) and could not be tested because of this.

** These dataframes contained features extracted from graphs inferred without the proper tuning of parameters of the `rLogSpecT` method.

† 435 edge weights (between 30 ICA components) and 130 remaining features.

N/A: not applicable; N/S: not specified.

D Formulas of the Performance Metrics

D.1 Confusion Matrix Terms

- **TP**: True Positives (correctly predicted ASD)
- **TN**: True Negatives (correctly predicted control)
- **FP**: False Positives (control predicted as ASD)
- **FN**: False Negatives (ASD predicted as control)

D.2 Accuracy

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{D.1}$$

D.3 Sensitivity

$$sensitivity = \frac{TP}{TP + FN} \tag{D.2}$$

D.4 Specificity

$$specificity = \frac{TN}{TN + FP} \tag{D.3}$$

D.5 Balanced Accuracy

$$balanced\ accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) = \frac{1}{2}\left(sensitivity + specificity\right) \tag{D.4}$$

D.6 AUROC

$$AUROC = \frac{1}{2}(sensitivity + specificity) \tag{D.5}$$

D.7 Precision

$$precision = \frac{TP}{TP + FP} \tag{D.6}$$

D.8 Recall

$$recall = \frac{TP}{TP + FN} \tag{D.7}$$

D.9 F1-score

$$f1\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \tag{D.8}$$

E Additional Figures

E.1 Connectome plots

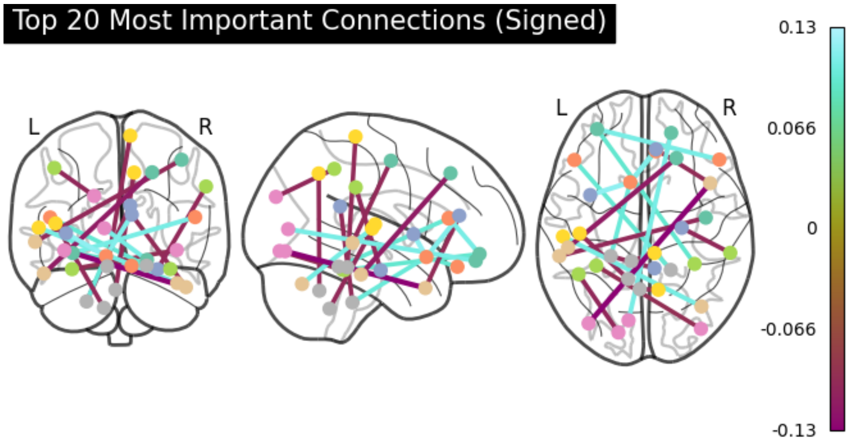


Figure E.1: Connectome visualisation for logistic regression, fold 1. Top 20 most important features are shown, weighted by importance score.

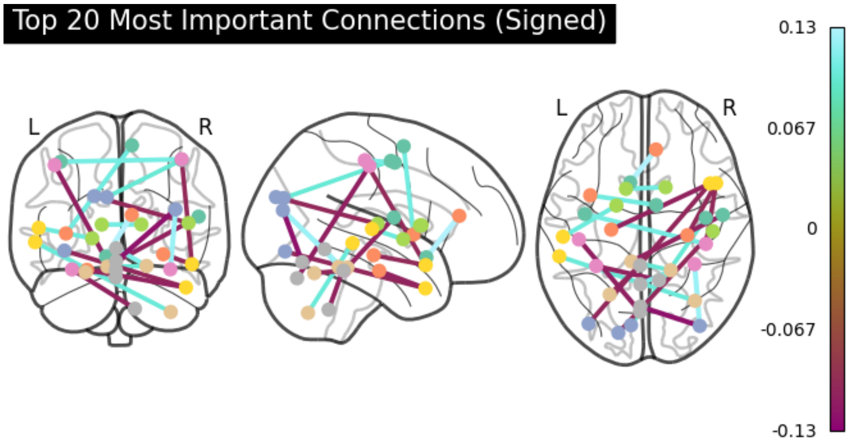


Figure E.2: Connectome visualisation for logistic regression, fold 2. Top 20 most important features are shown, weighted by importance score.

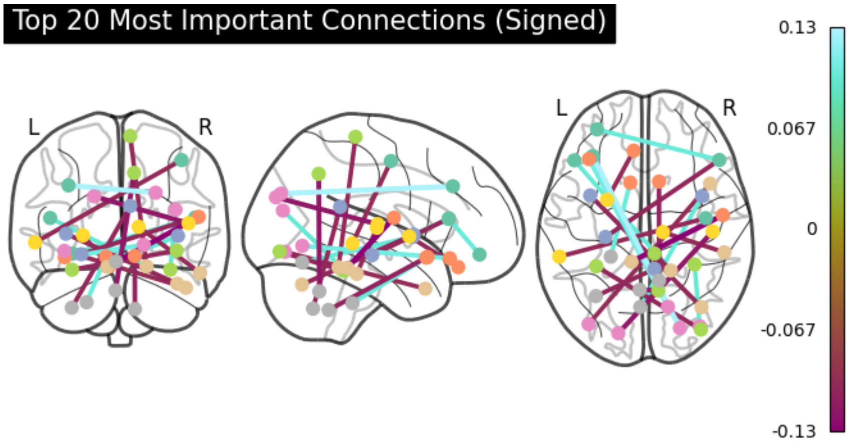


Figure E.3: Connectome visualisation for logistic regression, fold 3. Top 20 most important features are shown, weighted by importance score.

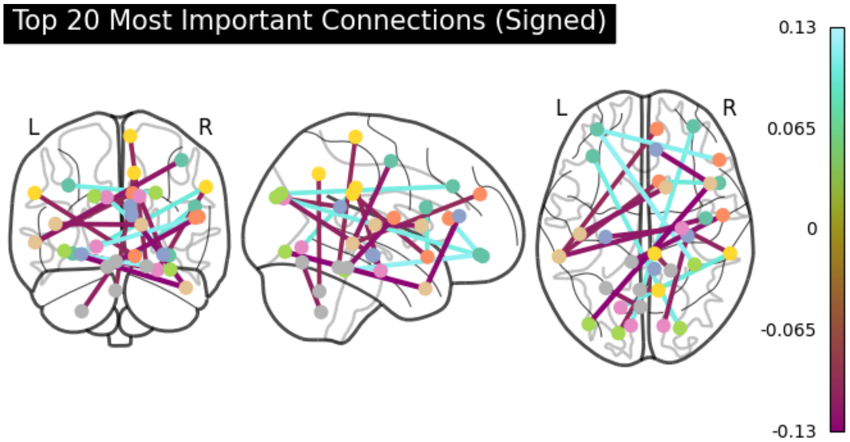


Figure E.4: Connectome visualisation for logistic regression, fold 4. Top 20 most important features are shown, weighted by importance score.

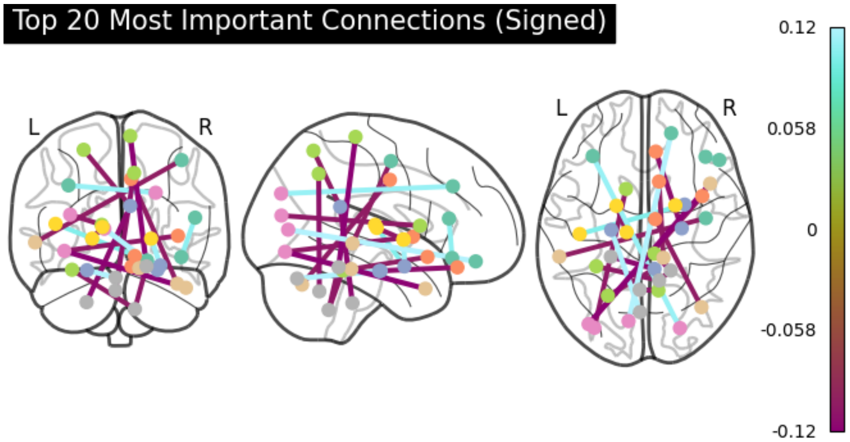


Figure E.5: Connectome visualisation for logistic regression, fold 5. Top 20 most important features are shown, weighted by importance score.

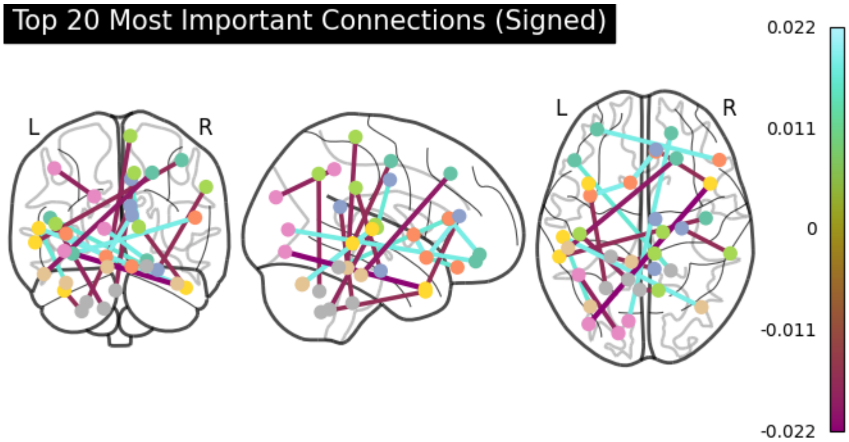


Figure E.6: Connectome visualisation for SVM, fold 1. Top 20 most important features are shown, weighted by importance score.

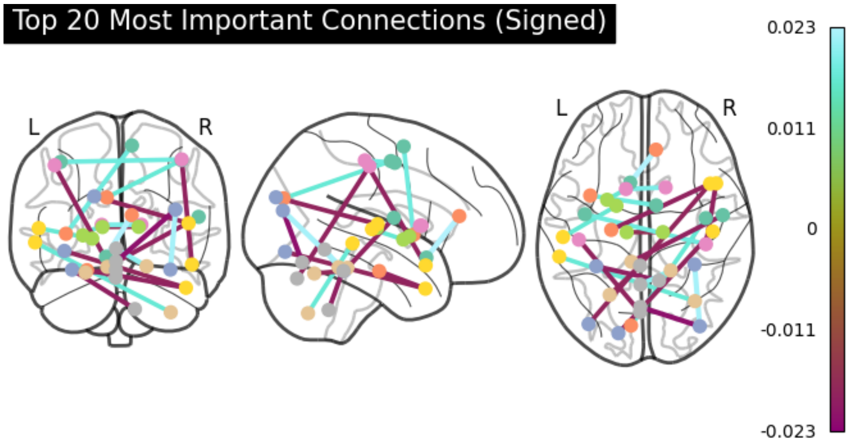


Figure E.7: Connectome visualisation for SVM, fold 2. Top 20 most important features are shown, weighted by importance score.

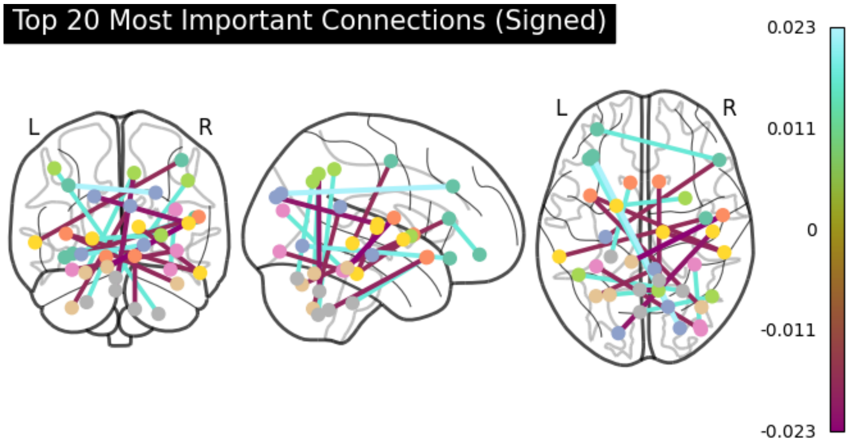


Figure E.8: Connectome visualisation for SVM, fold 3. Top 20 most important features are shown, weighted by importance score.

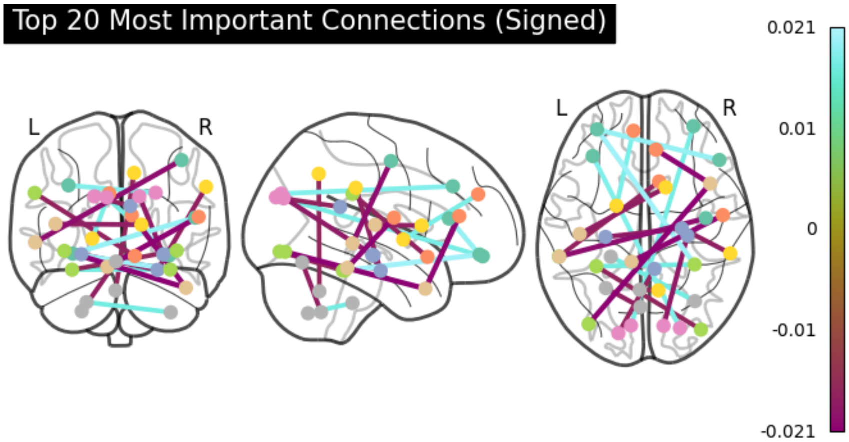


Figure E.9: Connectome visualisation for SVM, fold 4. Top 20 most important features are shown, weighted by importance score.

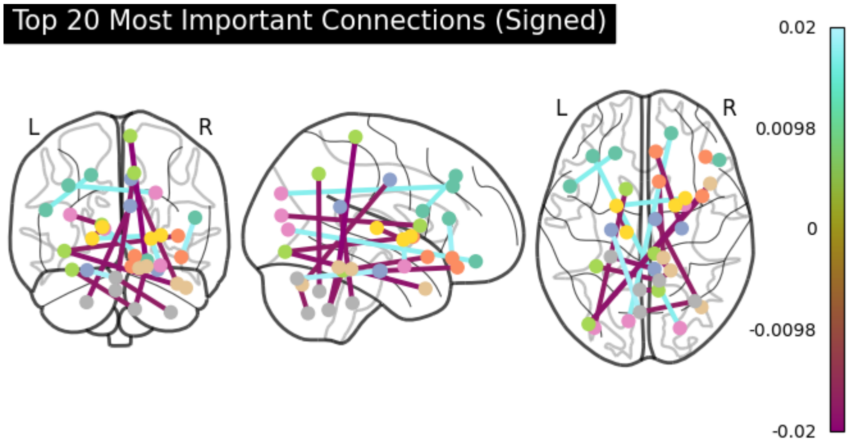


Figure E.10: Connectome visualisation for SVM, fold 5. Top 20 most important features are shown, weighted by importance score.

F Additional Tables

F.1 Multi-Site Classification on Pearson Correlation Features

F.1.1 Per-site Performances

Table F.1: The average performance across folds on raw combined data, evaluated per site. Comparison between **CALTECH** (left) and **KKI** (right).

	CALTECH					**KKI**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	33.3±57.7	60.0±52.9	56.7±40.4	55.6±9.6	46.7±5.8	48.0±50.2	58.3±37.3	51.9±36.3	50.0±13.8	53.2±17.6
SVM	33.3±57.7	60.0±52.9	56.7±40.4	55.6±9.6	46.7±5.8	56.0±51.8	55.0±38.9	58.6±40.5	51.9±13.6	55.5±17.0
DT	50.0±50.0	43.3±40.4	38.3±46.5	44.4±41.9	46.7±45.1	47.3±39.2	65.0±22.4	**72.5±25.3**	55.4±21.0	56.2±25.3
RF	**83.3±28.9**	**100.0±0.0**	**100.0±0.0**	**91.7±14.4**	**91.7±14.4**	32.0±46.0	**88.3±11.2**	60.6±31.2	**68.8±20.6**	60.2±25.9
MLP	33.3±57.7	60.0±52.9	68.3±38.8	55.6±9.6	46.7±5.8	65.3±40.9	58.3±40.8	52.8±39.6	59.6±23.6	61.8±24.1
LDA	66.7±57.7	60.0±52.9	68.3±38.8	61.1±19.2	63.3±23.1	**72.7±30.0**	56.7±38.4	70.3±28.1	59.9±19.6	**64.7±17.1**
KNN	50.0±50.0	36.7±32.1	40.0±17.3	50.0±0.0	43.3±11.5	48.7±36.6	53.3±36.1	52.1±35.5	48.5±28.5	51.0±31.3

Table F.2: The average performance across folds on raw combined data, evaluated per site. Comparison between **LEUVEN_1** (left) and **LEUVEN_2** (right).

	LEUVEN_1					**LEUVEN_2**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	35.7±38.0	76.7±22.4	**70.0±19.2**	54.8±13.5	56.2±16.6	**60.0±41.8**	65.3±40.9	78.3±18.3	69.9±21.5	62.7±26.4
SVM	37.3±44.4	76.7±22.4	64.3±26.3	54.8±22.1	57.0±23.2	**60.0±41.8**	75.3±23.3	80.8±18.5	**73.2±22.1**	67.7±25.8
DT	**84.3±15.1**	66.7±20.4	67.7±10.2	**71.2±10.5**	**75.5±12.1**	15.0±22.4	52.0±37.5	37.6±29.0	34.6±24.3	33.5±26.8
RF	5.0±11.2	**100.0±0.0**	59.8±36.0	56.9±17.5	52.5±5.6	15.0±33.5	**100.0±0.0**	79.3±21.8	67.0±10.4	57.5±16.8
MLP	46.3±38.5	81.7±17.1	69.3±25.0	64.3±22.6	64.0±25.7	50.0±50.0	76.7±22.4	**86.0±14.2**	71.0±17.9	63.3±20.9
LDA	37.0±43.0	56.7±36.5	53.0±28.2	51.9±10.2	46.8±17.8	35.0±33.5	54.7±44.1	53.7±15.3	50.6±22.1	44.8±21.1
KNN	10.7±15.3	83.3±23.6	50.7±18.5	48.1±14.3	47.0±17.6	50.0±0.0	86.7±29.8	74.8±22.8	70.5±18.4	**68.3±14.9**

Table F.3: The average performance across folds on raw combined data, evaluated per site. Comparison between **MAX_MUN** (left) and **NYU** (right).

	MAX_MUN					**NYU**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	41.7±20.4	54.5±23.1	61.4±18.6	48.3±17.0	48.1±13.6	65.4±19.4	72.4±11.8	74.7±10.7	69.0±13.8	**68.9±14.3**
SVM	41.7±20.4	62.4±20.6	59.7±16.4	53.4±17.9	52.0±15.8	62.1±14.6	65.8±12.6	69.7±6.8	63.9±9.2	63.9±9.4
DT	48.3±38.4	61.5±24.6	50.5±18.9	59.1±11.8	54.9±14.6	65.3±15.0	58.2±13.9	65.9±10.2	60.6±10.4	61.8±9.8
RF	21.7±21.7	59.8±36.1	53.3±20.1	48.8±19.6	40.7±8.5	43.5±12.4	**82.4±5.4**	72.0±12.0	65.4±9.4	63.0±8.6
MLP	**55.0±29.8**	63.5±22.1	64.8±14.2	57.0±12.2	59.3±9.1	**68.9±13.8**	63.6±10.3	75.3±11.1	66.0±4.4	66.3±4.8
LDA	48.3±29.1	**78.1±25.6**	**68.5±29.8**	**65.8±18.9**	**63.2±20.8**	65.1±6.6	72.4±6.5	**75.9±3.2**	**69.7±4.9**	68.8±4.1
KNN	40.0±25.3	46.9±15.4	43.9±13.1	44.9±11.2	43.4±11.9	44.0±25.1	67.5±17.9	64.6±10.6	58.8±9.7	55.7±12.5

Table F.4: The average performance across folds on raw combined data, evaluated per site. Comparison between **OHSU** (left) and **OLIN** (right).

	OHSU					**OLIN**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	37.5±28.5	20.8±25.0	32.3±37.3	32.1±23.7	29.2±25.0	68.7±18.8	66.7±23.6	78.2±13.6	63.0±8.3	67.7±7.8
SVM	25.0±31.9	27.1±35.6	23.3±27.9	32.1±23.7	26.0±22.1	68.7±18.8	56.7±14.9	72.0±15.9	60.1±7.0	62.7±5.6
DT	**50.0±40.8**	47.9±44.3	**49.7±40.5**	43.5±35.4	**49.0±38.2**	75.3±23.3	63.3±24.7	65.0±9.1	**68.4±10.8**	69.3±13.5
RF	25.0±21.5	33.3±23.6	38.2±20.1	32.7±14.6	29.2±17.3	68.7±18.8	**73.3±27.9**	81.6±19.0	67.0±10.9	**71.0±6.2**
MLP	33.3±30.4	31.2±37.5	39.6±36.9	36.9±21.8	32.3±22.7	75.3±23.3	63.3±24.7	**81.8±13.7**	66.8±15.6	69.3±13.5
LDA	33.3±30.4	**50.0±40.8**	26.0±32.3	**45.2±29.3**	41.7±32.6	62.0±24.7	70.0±29.8	72.0±15.9	65.8±12.4	66.0±13.8
KNN	37.5±47.9	35.4±29.2	36.8±37.5	44.0±28.3	36.5±31.8	**85.3±20.2**	56.7±27.9	69.8±30.2	68.1±22.5	**71.0±20.7**

Table F.5: The average performance across folds on raw combined data, evaluated per site. Comparison between **PITT** (left) and **SBL** (right).

	PITT					SBL				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	63.9±29.9	**91.0±12.4**	**82.0±20.0**	**76.5±18.2**	**77.4±17.4**	47.5±41.1	66.7±23.6	55.8±21.9	52.8±11.1	57.1±12.0
SVM	58.9±16.0	84.5±15.9	77.1±18.5	71.4±12.5	71.7±12.2	**52.5±41.1**	70.8±25.0	55.0±22.7	**57.3±15.0**	**61.7±13.5**
DT	59.4±38.6	68.0±12.5	61.3±29.2	60.6±21.9	63.7±23.2	25.0±28.9	64.6±17.2	45.3±18.7	45.5±19.3	44.8±15.7
RF	43.3±36.5	87.5±12.5	66.9±23.0	65.5±23.1	65.4±22.4	5.0±10.0	**85.4±17.2**	55.8±42.8	46.4±4.8	45.2±6.0
MLP	58.3±30.0	90.0±13.7	81.1±16.8	72.5±16.6	74.2±14.8	47.5±41.1	72.9±20.8	61.7±25.8	57.0±18.6	60.2±18.3
LDA	60.6±12.8	86.0±12.9	73.5±4.8	72.7±5.7	73.3±6.7	42.5±43.5	66.7±23.6	**66.5±14.0**	50.5±11.5	54.6±14.0
KNN	**67.2±27.7**	72.0±12.5	69.9±18.2	70.3±10.9	69.6±13.7	42.5±43.5	70.8±34.4	48.0±28.7	52.2±22.4	56.7±23.2

Table F.6: The average performance across folds on raw combined data, evaluated per site. Comparison between **SDSU** (left) and **STANFORD** (right).

	SDSU					STANFORD				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	33.3±47.1	78.3±15.8	70.3±27.7	64.3±24.0	55.8±30.2	55.8±30.2	25.0±28.9	**64.0±27.4**	46.1±17.3	40.4±16.0
SVM	33.3±47.1	78.3±15.8	61.9±43.2	64.3±24.0	55.8±30.2	55.8±30.2	50.0±40.8	**64.0±27.4**	52.4±10.1	52.9±9.5
DT	33.3±23.6	61.7±43.3	51.4±40.7	50.4±35.9	47.5±33.3	53.3±19.4	37.5±47.9	38.1±27.0	49.7±17.8	45.4±24.2
RF	**45.8±41.7**	**95.0±10.0**	**88.3±14.5**	**78.3±14.9**	**70.4±21.1**	43.3±41.6	**62.5±47.9**	40.5±30.9	56.5±12.0	52.9±9.5
MLP	33.3±47.1	65.0±3.3	67.8±28.2	55.7±16.5	49.2±24.4	60.8±28.3	31.2±37.5	46.2±32.5	53.9±19.8	46.0±20.2
LDA	41.7±28.9	65.0±27.4	52.5±14.5	57.6±18.5	53.3±21.9	**72.5±32.0**	37.5±47.9	59.6±33.8	**63.4±30.9**	**55.0±33.2**
KNN	**45.8±41.7**	48.3±17.5	42.9±35.2	48.9±24.8	47.1±28.8	64.2±26.3	25.0±50.0	49.3±15.4	47.6±21.7	44.6±19.9

Table F.7: The average performance across folds on raw combined data, evaluated per site. Comparison between **TRINITY** (left) and **UCLA_1** (right).

	TRINITY					UCLA_1				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	67.3±38.0	59.1±24.5	**80.7±25.0**	53.0±16.6	63.2±14.7	**70.3±18.9**	68.3±25.7	74.8±19.1	66.2±20.8	69.3±19.9
SVM	**74.0±33.2**	64.1±27.7	80.3±27.4	59.2±21.4	**69.0±14.7**	63.3±19.3	73.3±26.7	74.7±16.4	66.7±20.3	68.3±19.2
DT	43.0±33.0	**76.4±35.7**	62.8±24.1	51.7±18.5	59.7±8.7	49.3±24.9	69.1±25.2	57.9±21.5	54.8±18.6	59.2±18.5
RF	66.3±23.3	71.6±20.6	70.0±21.8	**63.2±14.9**	69.0±14.8	33.7±18.5	88.8±11.0	62.5±7.7	57.5±10.4	61.2±11.5
MLP	71.3±27.8	61.9±22.9	79.8±22.6	58.0±11.5	66.6±15.0	60.0±42.4	80.6±18.6	75.4±17.4	69.8±18.2	70.3±20.7
LDA	58.0±26.6	64.4±23.6	74.0±24.5	54.8±5.3	61.2±12.4	52.3±29.5	67.3±23.4	64.0±29.2	59.2±23.0	59.8±23.8
KNN	67.0±24.2	45.1±32.1	58.1±15.2	48.1±13.6	56.0±20.6	59.3±9.2	**93.5±9.3**	**86.2±15.0**	**74.0±12.2**	**76.4±9.0**

Table F.8: The average performance across folds on raw combined data, evaluated per site. Comparison between **UCLA_2** (left) and **UM_1** (right).

	UCLA_2					UM_1				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	**73.3±43.5**	40.0±41.8	76.7±43.5	59.6±36.7	56.7±37.0	50.4±29.9	62.6±12.0	64.8±16.2	55.6±14.9	56.5±12.9
SVM	**73.3±43.5**	45.0±44.7	73.3±43.5	62.5±38.3	59.2±38.9	57.1±16.6	55.1±11.2	61.7±17.7	55.9±11.2	56.1±10.7
DT	60.0±43.5	65.0±41.8	60.0±45.4	62.3±42.7	62.5±42.5	50.3±18.2	50.8±16.2	50.1±11.2	51.0±5.6	50.5±2.3
RF	53.3±36.1	80.0±44.7	68.3±43.5	63.8±38.3	66.7±38.6	47.1±18.1	70.1±11.8	65.6±9.6	60.6±9.2	58.6±11.0
MLP	70.0±44.7	50.0±50.0	66.7±42.5	61.0±43.7	60.0±45.4	64.7±33.6	71.2±9.3	**69.3±19.2**	65.4±13.4	67.9±13.6
LDA	**73.3±43.5**	50.0±50.0	**78.3±43.9**	63.6±37.8	61.7±38.9	56.7±29.9	**81.5±10.7**	69.1±23.6	**70.4±17.6**	**69.1±15.3**
KNN	50.0±37.3	**85.0±22.4**	75.8±30.7	**66.3±23.9**	**67.5±27.1**	**64.7±24.0**	59.8±14.3	68.1±15.1	60.8±15.8	62.3±14.4

Table F.9: The average performance across folds on raw combined data, evaluated per site. Comparison between **UM_2** (left) and **USM** (right).

	UM_2					**USM**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	54.2±8.3	70.4±24.7	68.1±18.2	63.6±17.8	62.3±15.9	84.2±16.8	50.3±33.2	68.2±8.4	**72.8±4.7**	67.3±10.4
SVM	**62.5±25.0**	76.7±29.1	69.8±21.3	69.9±23.7	69.6±23.7	71.8±19.2	57.0±40.0	69.3±9.9	67.8±10.9	64.4±15.0
DT	54.2±36.3	62.5±26.3	53.8±22.2	59.8±7.1	58.3±8.1	72.4±26.9	51.2±21.1	68.0±15.8	65.2±11.6	61.8±5.0
RF	31.2±23.9	**95.0±10.0**	68.8±23.5	68.3±11.2	63.1±13.7	69.8±26.3	**61.0±28.3**	**73.5±11.6**	66.7±8.1	65.4±3.6
MLP	**62.5±25.0**	90.0±20.0	**75.2±20.4**	**77.6±18.3**	**76.2±18.4**	80.2±14.3	42.3±28.8	67.8±18.5	67.6±10.0	61.3±12.8
LDA	58.3±28.9	75.4±17.5	74.7±17.5	68.3±11.2	66.9±12.0	74.9±15.4	60.8±17.4	71.3±14.0	70.0±14.7	**67.8±14.1**
KNN	47.9±17.2	57.1±37.0	56.2±23.5	53.1±28.3	52.5±26.3	**86.8±12.4**	41.2±28.9	72.1±14.8	71.1±11.3	64.0±15.3

Table F.10: The average performance across folds on raw combined data, evaluated on **YALE**.

	YALE				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	61.0±19.2	79.0±20.1	67.9±22.3	70.8±12.9	70.0±15.6
SVM	49.3±30.3	74.3±15.3	67.9±24.1	65.5±10.6	61.8±17.0
DT	37.0±17.8	63.7±25.3	55.9±11.8	50.7±13.9	50.3±14.3
RF	51.3±9.6	**93.0±11.0**	**84.3±12.2**	**71.6±8.3**	**72.2±6.2**
MLP	**66.0±18.9**	65.3±14.6	79.6±9.8	68.1±9.7	65.7±12.3
LDA	59.3±21.5	77.3±20.9	71.3±19.3	69.5±14.3	68.3±16.5
KNN	34.7±27.2	84.3±15.1	62.9±8.7	63.2±14.4	59.5±12.6

F.1.2 Per-age Performances

Table F.11: The average performance across folds on raw combined data, evaluated per age group. Comparison between **0-11** (left) and **12-18** (right).

	0-11					**12-18**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	57.8±13.3	55.9±9.3	61.7±10.1	55.8±8.4	56.8±9.3	61.1±6.8	69.2±8.6	72.9±3.6	65.0±1.2	65.1±1.1
SVM	52.4±6.0	57.7±9.8	59.8±8.7	55.2±7.9	55.0±7.8	61.9±7.0	69.1±7.5	71.2±3.8	65.5±1.0	65.5±1.2
DT	59.4±22.7	54.7±9.2	60.7±15.2	56.7±13.3	57.0±13.2	49.6±11.4	60.1±6.6	55.9±6.6	55.3±4.7	54.8±4.4
RF	42.9±10.8	**66.4±7.7**	59.5±10.2	56.3±5.0	54.7±5.6	44.0±11.9	**82.4±3.1**	70.7±6.4	64.4±5.7	63.2±6.0
MLP	**61.9±14.9**	54.4±9.0	**64.3±10.4**	57.0±7.2	58.2±6.9	**65.2±6.4**	69.1±8.4	**73.5±4.0**	**66.9±4.1**	**67.1±3.9**
LDA	61.3±13.5	58.7±9.9	63.6±12.6	**59.2±8.4**	**60.0±8.2**	60.5±7.6	68.6±6.1	70.0±4.8	64.6±5.8	64.5±6.0
KNN	52.7±14.2	57.0±6.1	58.6±8.9	53.6±5.2	54.8±5.5	60.3±10.2	66.7±5.6	67.3±5.4	63.3±4.2	63.5±4.3

Table F.12: The average performance across folds on raw combined data, evaluated per age group. Comparison between **19-30** (left) and **30+** (right).

	19-30					**30+**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	55.9±9.2	69.8±10.6	64.9±5.6	64.0±7.6	62.8±6.9	57.0±11.0	62.5±16.4	61.3±13.0	58.5±5.2	59.7±4.9
SVM	54.8±11.4	66.8±6.6	62.1±2.8	61.6±6.1	60.8±5.7	57.0±11.0	66.5±18.0	62.1±12.6	59.9±4.3	61.7±4.5
DT	**61.0±14.6**	58.1±13.2	58.9±9.2	59.5±11.2	59.5±10.9	48.0±18.2	51.2±18.8	51.1±9.2	48.8±12.3	49.6±12.5
RF	36.7±7.9	**83.9±6.1**	**68.5±10.4**	64.5±4.8	60.3±5.0	35.5±20.6	**81.3±11.9**	68.1±25.4	56.8±11.1	58.4±11.1
MLP	57.7±9.9	72.7±9.5	68.2±5.6	66.2±8.1	**65.2±7.7**	55.5±25.5	73.3±22.1	72.0±10.4	**62.1±4.3**	**64.4±5.2**
LDA	53.8±5.6	74.8±10.4	67.3±9.2	**66.4±7.0**	64.3±5.9	51.0±14.3	73.3±22.1	**72.6±11.0**	59.9±7.9	62.2±7.3
KNN	49.3±16.1	56.8±6.3	55.6±6.5	54.1±4.7	53.0±6.3	**61.5±15.4**	58.3±39.1	68.7±18.1	59.1±19.7	59.9±19.3

F.1.3 LOGO CV on Pearson Correlation

Table F.13: Sensitivity obtained by each classifier for each test site during LOGO CV on Pearson correlation features.

	Sensitivity [%]							**N**	
	LR	SVM	DT	RF	MLP	LDA	KNN	ASD	TC
CALTECH	40.0	40.0	20.0	20.0	**60.0**	20.0	40.0	5	10
CMU	**83.3**	**83.3**	**83.3**	33.3	**83.3**	50.0	16.7	6	5
KKI	**75.0**	**75.0**	41.7	41.7	**75.0**	**75.0**	**75.0**	12	21
LEUVEN_1	14.3	7.1	**42.9**	7.1	21.4	28.6	21.4	14	14
LEUVEN_2	58.3	**66.7**	58.3	16.7	50.0	50.0	**66.7**	12	16
MAXMUN	42.1	31.6	26.3	47.4	36.8	**57.9**	31.6	19	27
NYU	77.0	**100.0**	62.2	58.1	81.1	70.3	71.6	74	98
OHSU	50.0	33.3	**66.7**	58.3	58.3	16.7	41.7	12	13
OLIN	78.6	78.6	42.9	64.3	**85.7**	71.4	**85.7**	14	14
PITT	54.2	50.0	54.2	20.8	54.2	41.7	**58.3**	24	26
SBL	0.0	33.3	**41.7**	25.0	16.7	25.0	33.3	12	14
SDSU	**62.5**	**62.5**	37.5	50.0	50.0	**62.5**	**62.5**	8	19
STANFORD	**66.7**	**66.7**	58.3	**66.7**	**66.7**	**66.7**	58.3	12	13
TRINITY	68.4	**73.7**	57.9	47.4	**73.7**	47.4	63.2	19	25
UCLA_1	**59.5**	54.1	43.2	29.7	54.1	**59.5**	51.4	37	27
UCLA_2	81.8	81.8	81.8	63.6	81.8	**90.9**	63.6	11	10
UM_1	44.1	50.0	70.6	58.8	55.9	64.7	**76.5**	34	52
UM_2	**53.8**	46.2	23.1	46.2	**53.8**	38.5	46.2	13	21
USM	**72.1**	**72.1**	46.5	46.5	**72.1**	69.8	39.5	43	24
YALE	68.2	68.2	36.4	45.5	**72.7**	40.9	50.0	22	19
Mean±std	57.5±21.1	58.7±21.8	49.8±17.3	42.4±17.1	60.2±18.9	52.4±19.5	52.7±18.3		

Table F.14: Specificity obtained by each classifier for each test site during LOGO CV on Pearson Correlation features.

	Specificity [%]							**N**	
	LR	SVM	DT	RF	MLP	LDA	KNN	ASD	TC
CALTECH	50.0	50.0	70.0	**90.0**	70.0	60.0	40.0	5	10
CMU	80.0	80.0	60.0	**100.0**	**100.0**	**100.0**	60.0	6	5
KKI	57.1	52.4	**85.7**	81.0	57.1	57.1	57.1	12	21
LEUVEN_1	92.9	92.9	57.1	**100.0**	85.7	85.7	92.9	14	14
LEUVEN_2	43.8	37.5	56.2	**93.8**	50.0	62.5	81.2	12	16
MAXMUN	55.6	55.6	**85.2**	66.7	59.3	63.0	63.0	19	27
NYU	52.0	7.1	55.1	**68.4**	37.8	49.0	53.1	74	98
OHSU	38.5	38.5	46.2	**69.2**	46.2	61.5	46.2	12	13
OLIN	57.1	57.1	**71.4**	64.3	57.1	64.3	57.1	14	14
PITT	**92.3**	88.5	61.5	84.6	**92.3**	88.5	69.2	24	26
SBL	71.4	85.7	28.6	**92.9**	78.6	85.7	50.0	12	14
SDSU	**84.2**	73.7	52.6	73.7	78.9	57.9	57.9	8	19
STANFORD	46.2	46.2	**69.2**	**69.2**	46.2	30.8	15.4	12	13
TRINITY	44.0	**72.0**	56.0	52.0	52.0	64.0	48.0	19	25
UCLA_1	**85.2**	**85.2**	77.8	**85.2**	81.5	66.7	**85.2**	37	27
UCLA_2	60.0	60.0	**90.0**	70.0	50.0	50.0	70.0	11	10
UM_1	71.2	**73.1**	69.2	67.3	61.5	53.8	51.9	34	52
UM_2	71.4	**90.5**	81.0	85.7	**90.5**	**90.5**	52.4	13	21
USM	70.8	75.0	75.0	**83.3**	79.2	66.7	70.8	43	24
YALE	78.9	78.9	57.9	73.7	78.9	**84.2**	73.7	22	19
Mean±std	65.1±16.5	65.0±21.5	65.3±14.7	78.5±12.7	67.6±17.6	67.1±16.6	59.8±16.9		

Table F.15: Auroc obtained by each classifier for each test site during LOGO CV on Pearson Correlation features.

	AUROC [%]							**N**	
	LR	SVM	DT	RF	MLP	LDA	KNN	ASD	TC
CALTECH	44.0	46.0	53.0	**62.0**	58.0	58.0	40.0	5	10
CMU	**93.3**	**93.3**	71.7	**93.3**	86.7	86.7	48.3	6	5
KKI	62.3	68.3	64.3	**71.8**	70.6	69.8	55.2	12	21
LEUVEN_1	59.2	**69.9**	57.4	64.3	67.9	49.0	54.3	14	14
LEUVEN_2	55.2	56.2	53.9	65.6	60.9	62.0	**73.2**	12	16
MAXMUN	52.8	42.7	**59.0**	54.4	48.0	56.3	42.8	19	27
NYU	**71.4**	68.5	58.9	69.1	71.1	70.9	65.8	74	98
OHSU	46.2	38.5	**70.8**	52.9	48.7	48.1	39.4	12	13
OLIN	75.5	73.5	60.5	67.1	**80.6**	74.0	71.4	14	14
PITT	**76.1**	75.6	57.0	65.5	71.2	67.3	60.1	24	26
SBL	57.1	**73.2**	31.0	65.2	60.1	57.7	44.9	12	14
SDSU	73.0	**78.0**	42.4	67.8	75.0	60.5	59.9	8	19
STANFORD	62.2	62.8	**73.7**	63.5	63.5	48.7	49.7	12	13
TRINITY	62.3	**70.1**	61.8	52.9	68.6	62.7	61.3	19	25
UCLA_1	74.7	74.8	60.6	68.7	70.4	65.0	**80.4**	37	27
UCLA_2	78.2	78.2	**88.6**	71.8	82.7	75.5	81.4	11	10
UM_1	71.3	71.5	69.9	**74.2**	63.3	69.4	70.1	34	52
UM_2	71.1	72.9	56.8	68.5	78.4	**83.5**	54.6	13	21
USM	75.2	75.1	66.4	76.4	**79.2**	72.6	59.6	43	24
YALE	**82.8**	82.3	51.0	67.8	78.5	77.8	66.0	22	19
Mean±std	67.2±12.2	68.6±13.2	60.4±11.8	67.1±8.7	69.2±10.4	65.8±10.8	58.9±12.2		

Table F.16: Accuracy obtained by each classifier for each test site during LOGO CV on Pearson correlation features.

	Accuracy [%]							**N**	
	LR	SVM	DT	RF	MLP	LDA	KNN	ASD	TC
CALTECH	46.7	46.7	53.3	**66.7**	**66.7**	46.7	40.0	5	10
CMU	81.8	81.8	72.7	63.6	**90.9**	72.7	36.4	6	5
KKI	63.6	60.6	**69.7**	66.7	63.6	63.6	63.6	12	21
LEUVEN_1	53.6	50.0	50.0	53.6	53.6	**57.1**	**57.1**	14	14
LEUVEN_2	50.0	50.0	57.1	60.7	50.0	57.1	**75.0**	12	16
MAXMUN	50.0	45.7	**60.9**	58.7	50.0	**60.9**	50.0	19	27
NYU	62.8	47.1	58.1	**64.0**	56.4	58.1	61.0	74	98
OHSU	44.0	36.0	56.0	**64.0**	52.0	40.0	44.0	12	13
OLIN	67.9	67.9	57.1	64.3	**71.4**	67.9	**71.4**	14	14
PITT	**74.0**	70.0	58.0	54.0	**74.0**	66.0	64.0	24	26
SBL	38.5	**61.5**	34.6	**61.5**	50.0	57.7	42.3	12	14
SDSU	**77.8**	70.4	48.1	66.7	70.4	59.3	59.3	8	19
STANFORD	56.0	56.0	64.0	**68.0**	56.0	48.0	36.0	12	13
TRINITY	54.5	**72.7**	56.8	50.0	61.4	56.8	54.5	19	25
UCLA_1	**70.3**	67.2	57.8	53.1	65.6	62.5	65.6	37	27
UCLA_2	71.4	71.4	**85.7**	66.7	66.7	71.4	66.7	11	10
UM_1	60.5	64.0	**69.8**	64.0	59.3	58.1	61.6	34	52
UM_2	64.7	73.5	58.8	70.6	**76.5**	70.6	50.0	13	21
USM	71.6	73.1	56.7	59.7	**74.6**	68.7	50.7	43	24
YALE	73.2	73.2	46.3	58.5	**75.6**	61.0	61.0	22	19
Mean±std	61.6±11.8	61.9±12.1	58.6±10.4	61.7±5.5	64.2±10.8	60.2±8.2	55.5±11.2		

Table F.17: Top discriminative brain connectivity features selected across folds and models. Each checkmark indicates that the coresponding feature was among the top 20 most important features for a given fold and model.

Feature name	LR-F1	LR-F2	LR-F3	LR-F4	LR-F5	SVM-F1	SVM-F2	SVM-F3	SVM-F4	SVM-F5
fc_5301_8212	✓	✓	✓	✓	✓	✓	✓		✓	✓
fc_2501_3001	✓					✓				
fc_2102_8301	✓					✓				
fc_2501_4002	✓	✓				✓	✓			
fc_2211_2312	✓		✓	✓		✓		✓	✓	
fc_2112_5002	✓									
fc_8111_9002	✓					✓				
fc_5101_6201	✓					✓				
fc_4112_6212	✓			✓		✓			✓	

Continued on next page.

Table F.17 (continued from previous page)

Feature name	LR-F1	LR-F2	LR-F3	LR-F4	LR-F5	SVM-F1	SVM-F2	SVM-F3	SVM-F4	SVM-F5
fc_8101_9071	✓					✓				
fc_2002_8201	✓		✓	✓	✓	✓		✓	✓	
fc_6402_9021	✓					✓				
fc_2501_9081	✓		✓					✓		
fc_6302_9160	✓		✓	✓	✓	✓		✓	✓	✓
fc_4022_9081	✓				✓	✓				✓
fc_2311_9032	✓		✓			✓				
fc_5302_5401	✓		✓					✓	✓	
fc_2501_2702	✓					✓				
fc_2211_5402	✓			✓					✓	
fc_5202_9130		✓					✓			
fc_5202_5402		✓	✓				✓	✓		
fc_9032_9120		✓					✓			
fc_6001_9140		✓					✓			
fc_4111_8212		✓					✓			
fc_2332_9021		✓	✓	✓			✓	✓	✓	
fc_5401_9072		✓			✓		✓			✓
fc_2001_6002		✓					✓			
fc_7001_7002		✓					✓			
fc_9041_9100		✓					✓			
fc_2402_3001		✓					✓			
fc_7021_8111		✓					✓			
fc_5101_8102		✓	✓				✓	✓		
fc_6002_8122		✓					✓			
fc_4102_8122		✓								
fc_8201_9052		✓					✓			
fc_2002_5011		✓					✓			
fc_9021_9032		✓								
fc_2201_5102			✓	✓	✓			✓	✓	✓
fc_2502_9072			✓					✓		
fc_2321_5022			✓					✓		
fc_2312_4101			✓					✓		
fc_3001_9002			✓					✓		
fc_4022_9051			✓					✓		
fc_7102_8302			✓					✓		
fc_6402_9032			✓	✓	✓					✓
fc_2701_7011			✓							
fc_3002_9130			✓							
fc_4101_4112				✓					✓	
fc_7002_8101				✓					✓	
fc_4102_5012				✓					✓	
fc_9061_9130				✓					✓	
fc_2212_5011				✓					✓	
fc_2502_6211				✓					✓	
fc_4022_5101				✓					✓	
fc_2602_8201				✓						
fc_2302_2502				✓						
fc_5021_6212				✓						
fc_2112_5001					✓	✓				✓
fc_4111_9140					✓					✓
fc_2702_4112					✓					✓
fc_4202_7022					✓					✓
fc_7021_7101					✓					✓

Continued on next page.

Table F.17 (continued from previous page)

Feature name	LR-F1	LR-F2	LR-F3	LR-F4	LR-F5	SVM-F1	SVM-F2	SVM-F3	SVM-F4	SVM-F5
fc_2312_2322					✓					✓
fc_5201_7001					✓					✓
fc_9032_9072					✓					✓
fc_3002_5301					✓					✓
fc_4022_6101					✓					
fc_4012_9002					✓					✓
fc_2502_9022					✓					✓
fc_4202_8101					✓					
fc_4012_7102						✓				
fc_8111_9001						✓				
fc_8211_9061						✓				
fc_7001_7011							✓			
fc_7101_7102							✓			
fc_7012_7021								✓		✓
fc_3001_8302								✓		
fc_6201_9062								✓		
fc_6222_9140								✓		
fc_6302_9041								✓		
fc_4002_8212									✓	
fc_2601_7021									✓	
fc_9052_9081									✓	
fc_5401_5402									✓	
fc_9052_9140										✓
fc_2101_2301										✓

F.2 Multi-Site Classification on Graph Features

F.2.1 Average Performances of All Received Dataframes

Table F.18: Multisite classification performance on DF1.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	45.9% ± 4.7%	55.3% ± 3.3%	**52.5% ± 3.6%**	51.0% ± 2.3%	50.6% ± 2.3%
SVM	**47.1% ± 4.6%**	55.5% ± 4.4%	49.9% ± 4.6%	51.7% ± 3.3%	51.3% ± 3.3%
DT	35.5% ± 22.1%	61.6% ± 23.6%	49.3% ± 3.4%	*49.5% ± 3.5%*	*48.5% ± 2.7%*
RF	*32.5% ± 5.5%*	**67.1% ± 6.1%**	51.3% ± 3.5%	51.1% ± 3.3%	49.8% ± 3.3%
MLP	44.9% ± 3.4%	*55.1% ± 11.7%*	50.5% ± 4.5%	50.4% ± 5.2%	50.0% ± 4.6%
LDA	46.9% ± 5.3%	56.0% ± 3.4%	52.4% ± 3.3%	**51.8% ± 3.0%**	**51.4% ± 3.1%**
KNN	33.0% ± 10.6%	66.7% ± 7.7%	*48.6% ± 6.6%*	51.1% ± 5.3%	49.8% ± 5.5%

Table F.19: Multisite classification performance on DF2.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	**45.7% ± 6.2%**	54.1% ± 7.9%	51.2% ± 4.0%	50.2% ± 5.0%	49.9% ± 4.9%
SVM	45.2% ± 8.7%	*52.4% ± 8.3%*	50.3% ± 4.5%	*49.0% ± 6.3%*	*48.8% ± 6.3%*
DT	45.4% ± 8.5%	55.3% ± 9.9%	49.6% ± 3.6%	50.7% ± 3.2%	50.4% ± 2.9%
RF	34.0% ± 5.4%	**65.4% ± 4.2%**	50.5% ± 3.3%	50.9% ± 2.4%	49.7% ± 2.5%
MLP	44.0% ± 8.6%	*62.4% ± 2.6%*	**53.0% ± 2.5%**	**53.8% ± 3.1%**	**53.2% ± 3.5%**
LDA	44.2% ± 4.5%	53.6% ± 6.9%	50.2% ± 3.7%	49.2% ± 4.6%	48.9% ± 4.4%
KNN	*39.7% ± 12.2%*	58.2% ± 12.7%	*49.3% ± 2.1%*	49.6% ± 1.3%	49.0% ± 0.8%

Table F.20: Multisite classification performance on DF3.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	45.9% ± 4.7%	*55.3% ± 3.3%*	**52.5% ± 3.6%**	51.0% ± 2.3%	50.6% ± 2.3%
SVM	**47.1% ± 4.6%**	55.5% ± 4.4%	50.0% ± 4.6%	*51.7% ± 3.3%*	51.3% ± 3.3%
DT	39.5% ± 23.1%	57.9% ± 25.7%	50.2% ± 3.9%	*49.4% ± 4.4%*	*48.7% ± 3.4%*
RF	33.5% ± 3.6%	**69.0% ± 6.2%**	51.4% ± 3.0%	**52.6% ± 2.9%**	51.2% ± 2.7%
MLP	44.7% ± 6.8%	55.8% ± 8.4%	50.8% ± 5.7%	50.6% ± 7.3%	50.2% ± 7.3%
LDA	46.9% ± 5.3%	56.0% ± 3.4%	52.4% ± 3.3%	51.8% ± 3.0%	**51.4% ± 3.1%**
KNN	*33.0% ± 10.6%*	66.7% ± 7.7%	*48.6% ± 6.6%*	51.1% ± 5.3%	49.8% ± 5.5%

Table F.21: Multisite classification performance on DF4.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	45.0% ± 8.4%	55.7% ± 4.1%	**51.5% ± 3.9%**	50.7% ± 2.7%	50.3% ± 3.0%
SVM	**46.2% ± 9.3%**	*54.0% ± 6.1%*	50.5% ± 4.6%	50.4% ± 3.9%	50.1% ± 4.1%
DT	41.7% ± 13.8%	61.5% ± 18.3%	51.3% ± 5.7%	**52.2% ± 6.7%**	**51.6% ± 6.2%**
RF	*34.1% ± 6.2%*	**65.0% ± 5.4%**	51.0% ± 4.5%	50.6% ± 2.5%	49.6% ± 2.6%
MLP	37.6% ± 8.9%	62.5% ± 12.4%	50.3% ± 3.4%	50.9% ± 4.9%	50.1% ± 4.5%
LDA	43.2% ± 8.0%	54.8% ± 5.0%	50.9% ± 3.7%	*49.4% ± 3.2%*	49.0% ± 3.3%
KNN	34.6% ± 10.5%	62.8% ± 10.5%	*47.6% ± 3.0%*	49.6% ± 1.7%	*48.7% ± 1.5%*

Table F.22: Multisite classification performance on DF5.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	54.8% ± 2.6%	63.7% ± 1.7%	63.2% ± 4.5%	59.6% ± 1.5%	59.3% ± 1.5%
SVM	51.1% ± 5.1%	61.1% ± 3.5%	60.0% ± 5.0%	56.5% ± 3.2%	56.1% ± 3.3%
DT	*38.0% ± 7.9%*	66.5% ± 7.2%	*55.2% ± 2.7%*	*53.3% ± 2.6%*	*52.2% ± 2.6%*
RF	39.5% ± 4.0%	**72.0% ± 3.8%**	58.8% ± 1.8%	57.0% ± 2.8%	55.7% ± 2.8%
MLP	55.4% ± 5.2%	65.2% ± 5.0%	63.4% ± 3.8%	60.6% ± 3.6%	60.3% ± 3.6%
LDA	**57.6% ± 3.6%**	63.9% ± 2.6%	**64.2% ± 4.4%**	**61.0% ± 1.6%**	**60.7% ± 1.7%**
KNN	48.4% ± 5.3%	*57.9% ± 5.6%*	*55.2% ± 2.3%*	53.5% ± 2.4%	53.1% ± 2.3%

Table F.23: Multisite classification performance on DF6.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	55.1% ± 2.9%	60.9% ± 4.6%	60.8% ± 4.9%	58.2% ± 3.6%	58.0% ± 3.5%
SVM	**56.6% ± 6.4%**	61.8% ± 3.3%	**62.5% ± 4.8%**	**59.4% ± 4.6%**	**59.2% ± 4.7%**
DT	44.2% ± 12.7%	64.1% ± 10.5%	*52.7% ± 2.9%*	*54.9% ± 1.5%*	*54.1% ± 1.9%*
RF	*43.7% ± 7.4%*	**69.6% ± 3.9%**	61.2% ± 3.0%	57.6% ± 3.9%	56.7% ± 4.1%
MLP	54.8% ± 9.6%	62.2% ± 4.8%	61.5% ± 7.4%	58.8% ± 6.3%	58.5% ± 6.5%
LDA	53.4% ± 4.7%	*59.8% ± 4.0%*	60.1% ± 5.3%	56.8% ± 3.8%	56.6% ± 3.8%
KNN	48.4% ± 6.0%	64.3% ± 5.2%	58.6% ± 2.6%	56.9% ± 2.6%	56.3% ± 2.7%

Table F.24: Multisite classification performance on DF7.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	*31.0% ± 7.4%*	73.1% ± 4.3%	53.5% ± 3.6%	53.6% ± 2.7%	52.1% ± 3.0%
SVM	31.5% ± 8.2%	**75.2% ± 3.7%**	**53.9% ± 3.0%**	**55.0% ± 2.9%**	**53.9% ± 3.0%**
DT	33.5% ± 5.4%	68.4% ± 6.1%	51.7% ± 4.7%	52.2% ± 3.2%	50.9% ± 3.2%
RF	31.5% ± 4.2%	64.3% ± 6.3%	*46.5% ± 5.9%*	49.1% ± 4.0%	47.9% ± 4.0%
MLP	34.0% ± 14.9%	65.4% ± 15.8%	49.8% ± 1.4%	50.9% ± 1.6%	49.7% ± 0.6%
LDA	31.5% ± 4.2%	71.6% ± 6.8%	52.6% ± 4.6%	53.0% ± 4.1%	51.6% ± 4.0%
KNN	**51.4% ± 5.8%**	*46.2% ± 5.0%*	47.9% ± 5.1%	*48.6% ± 4.5%*	*48.8% ± 4.5%*

Table F.25: Multisite classification performance on DF8.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	44.2% ± 2.9%	56.2% ± 2.0%	51.3% ± 3.0%	50.6% ± 2.2%	*50.2% ± 2.2%*
SVM	45.9% ± 2.9%	54.5% ± 2.8%	*49.4% ± 1.2%*	50.5% ± 2.4%	*50.2% ± 2.4%*
DT	51.6% ± 23.8%	*45.7% ± 24.6%*	50.5% ± 3.3%	*48.4% ± 2.6%*	50.5% ± 3.3%
RF	*30.8% ± 2.3%*	**74.1% ± 2.8%**	**52.0% ± 2.0%**	**54.1% ± 1.4%**	**52.5% ± 1.4%**
MLP	42.7% ± 6.2%	58.1% ± 6.3%	51.3% ± 3.7%	51.0% ± 3.9%	50.4% ± 3.9%
LDA	45.4% ± 3.1%	56.6% ± 4.8%	51.0% ± 2.9%	51.4% ± 3.5%	51.0% ± 3.4%
KNN	**53.9% ± 7.0%**	48.7% ± 7.4%	**52.0% ± 4.0%**	51.1% ± 4.1%	51.3% ± 4.1%

Table F.26: Multisite classification performance on DF9.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	43.2% ± 7.3%	58.8% ± 8.1%	50.4% ± 4.5%	51.5% ± 5.9%	51.0% ± 5.9%
SVM	42.4% ± 6.5%	59.8% ± 5.5%	50.7% ± 3.8%	51.8% ± 4.1%	51.1% ± 4.2%
DT	**53.1% ± 13.3%**	48.7% ± 14.6%	**52.7% ± 4.6%**	*50.7% ± 4.7%*	50.9% ± 4.4%
RF	*33.3% ± 3.5%*	**68.4% ± 4.2%**	50.4% ± 2.8%	**52.1% ± 2.8%**	50.8% ± 2.7%
MLP	43.4% ± 8.7%	*53.6% ± 8.7%*	*48.4% ± 3.1%*	48.9% ± 2.8%	*48.5% ± 2.7%*
LDA	44.7% ± 7.2%	58.6% ± 5.2%	49.9% ± 3.5%	**52.1% ± 3.5%**	**51.6% ± 3.7%**
KNN	44.7% ± 5.4%	56.0% ± 6.0%	51.2% ± 4.3%	*50.7% ± 4.7%*	50.3% ± 4.7%

Table F.27: Multisite classification performance on DF10.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	39.4% ± 7.1%	57.7% ± 7.2%	47.0% ± 3.0%	*49.3% ± 1.8%*	*48.6% ± 1.8%*
SVM	40.7% ± 6.8%	60.2% ± 6.8%	49.6% ± 3.5%	51.2% ± 3.4%	50.5% ± 3.4%
DT	43.7% ± 7.3%	*54.5% ± 6.0%*	49.6% ± 4.6%	49.5% ± 3.1%	49.1% ± 3.2%
RF	*34.2% ± 6.8%*	**69.9% ± 3.3%**	51.8% ± 3.2%	53.4% ± 3.3%	52.0% ± 3.5%
MLP	**46.7% ± 7.8%**	61.3% ± 4.5%	**54.1% ± 4.5%**	**54.5% ± 2.2%**	**54.0% ± 2.5%**
LDA	42.9% ± 6.7%	56.0% ± 7.3%	*46.5% ± 3.7%*	49.9% ± 2.6%	49.4% ± 2.4%
KNN	40.2% ± 5.2%	57.1% ± 4.6%	48.3% ± 3.0%	*49.3% ± 3.1%*	*48.6% ± 3.2%*

Table F.28: Multisite classification performance on DF11.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	43.7% ± 4.4%	59.4% ± 9.5%	**55.3% ± 7.4%**	52.1% ± 6.3%	51.5% ± 6.1%
SVM	41.4% ± 1.9%	62.4% ± 6.8%	*45.8% ± 7.0%*	**52.7% ± 3.7%**	**51.9% ± 3.5%**
DT	42.7% ± 4.4%	60.1% ± 7.6%	50.4% ± 2.4%	52.0% ± 2.4%	51.4% ± 2.1%
RF	*31.5% ± 5.6%*	**66.9% ± 7.5%**	50.6% ± 3.4%	50.5% ± 2.9%	49.2% ± 2.7%
MLP	40.6% ± 14.7%	59.2% ± 10.7%	51.5% ± 7.8%	50.6% ± 6.2%	49.9% ± 6.4%
LDA	**44.4% ± 4.6%**	57.7% ± 5.3%	51.3% ± 6.0%	51.5% ± 4.2%	51.1% ± 4.2%
KNN	*40.9% ± 6.9%*	*56.6% ± 4.9%*	49.1% ± 4.5%	*49.4% ± 3.5%*	*48.8% ± 3.6%*

Table F.29: Multisite classification performance on DF12.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	40.5% ± 6.9%	61.3% ± 7.3%	**50.6% ± 1.7%**	51.7% ± 2.6%	50.9% ± 2.5%
SVM	41.3% ± 7.6%	62.2% ± 6.0%	49.4% ± 2.5%	**52.5% ± 3.3%**	**51.8% ± 3.3%**
DT	43.6% ± 5.0%	52.0% ± 5.5%	*46.9% ± 3.6%*	*48.1% ± 3.3%*	*47.8% ± 3.3%*
RF	*29.9% ± 7.4%*	**66.9% ± 3.6%**	49.4% ± 1.9%	49.8% ± 2.3%	48.4% ± 2.5%
MLP	**42.3% ± 14.9%**	56.2% ± 11.8%	48.2% ± 4.9%	49.8% ± 4.4%	49.3% ± 4.6%
LDA	40.5% ± 4.5%	*55.8% ± 4.9%*	47.3% ± 2.8%	48.7% ± 1.8%	48.1% ± 1.7%
KNN	39.2% ± 4.4%	60.9% ± 6.5%	49.3% ± 3.2%	50.8% ± 5.4%	50.0% ± 5.3%

Table F.30: Multisite classification performance on DF13.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	39.7% ± 3.8%	59.8% ± 4.1%	49.4% ± 4.0%	50.5% ± 2.9%	49.8% ± 2.9%
SVM	38.0% ± 2.9%	57.9% ± 2.5%	**51.6% ± 4.1%**	48.7% ± 1.7%	47.9% ± 1.8%
DT	**48.9% ± 8.0%**	*50.9% ± 9.2%*	51.0% ± 3.7%	49.9% ± 2.5%	**49.9% ± 2.3%**
RF	*30.0% ± 2.9%*	**68.6% ± 4.4%**	48.7% ± 3.1%	**50.7% ± 2.5%**	49.3% ± 2.4%
MLP	39.9% ± 7.3%	58.7% ± 12.3%	49.4% ± 4.6%	50.1% ± 5.2%	49.3% ± 4.7%
LDA	40.7% ± 3.7%	53.9% ± 2.4%	*47.5% ± 3.2%*	*47.8% ± 2.0%*	*47.3% ± 2.0%*
KNN	35.0% ± 5.2%	59.6% ± 4.2%	48.2% ± 2.1%	48.2% ± 2.9%	*47.3% ± 2.9%*

Table F.31: Multisite classification performance on DF14.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	39.0% ± 6.5%	61.9% ± 7.7%	52.0% ± 4.0%	51.3% ± 3.5%	50.4% ± 3.3%
SVM	37.5% ± 5.0%	62.8% ± 9.3%	47.5% ± 2.4%	51.0% ± 4.8%	50.1% ± 4.5%
DT	38.9% ± 20.4%	63.4% ± 22.6%	52.1% ± 3.7%	52.1% ± 4.2%	51.2% ± 3.4%
RF	*32.0% ± 2.4%*	**63.8% ± 6.4%**	*45.8% ± 3.0%*	*49.1% ± 4.0%*	*47.9% ± 3.8%*
MLP	**50.7% ± 10.8%**	56.8% ± 8.3%	**52.2% ± 4.6%**	**53.9% ± 4.0%**	**53.7% ± 4.2%**
LDA	45.2% ± 6.1%	58.7% ± 6.1%	50.9% ± 3.0%	52.4% ± 4.1%	51.9% ± 4.0%
KNN	47.9% ± 5.7%	*52.7% ± 8.4%*	51.0% ± 5.7%	50.5% ± 5.6%	50.3% ± 5.5%

Table F.32: Multisite classification performance on DF15.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	33.4% ± 7.3%	66.5% ± 8.9%	51.8% ± 3.3%	51.2% ± 3.0%	50.0% ± 2.7%
SVM	*13.2% ± 2.3%*	**87.2% ± 3.7%**	49.9% ± 1.8%	52.9% ± 1.3%	50.2% ± 1.1%
DT	38.1% ± 6.3%	58.5% ± 9.6%	*48.4% ± 6.8%*	*49.0% ± 5.1%*	*48.3% ± 4.8%*
RF	43.6% ± 8.2%	64.2% ± 6.3%	**55.5% ± 3.4%**	**54.6% ± 2.5%**	**53.9% ± 2.7%**
MLP	42.7% ± 11.2%	57.8% ± 13.1%	51.0% ± 3.7%	50.8% ± 4.6%	50.3% ± 3.7%
LDA	33.7% ± 5.9%	67.3% ± 6.1%	51.2% ± 3.6%	51.7% ± 2.9%	50.5% ± 2.9%
KNN	**47.4% ± 12.1%**	*57.6% ± 4.0%*	52.6% ± 5.9%	52.9% ± 5.4%	52.5% ± 5.8%

Table F.33: Multisite classification performance on DF22.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	36.4% ± 7.1%	62.3% ± 4.1%	52.2% ± 3.6%	50.3% ± 2.6%	52.2% ± 3.6%
SVM	*28.2% ± 7.1%*	**76.8% ± 5.9%**	*48.9% ± 3.9%*	**54.3% ± 5.2%**	52.5% ± 5.3%
DT	39.2% ± 12.9%	64.7% ± 9.4%	50.9% ± 6.6%	52.9% ± 5.4%	52.0% ± 5.6%
RF	39.7% ± 9.4%	66.1% ± 5.8%	**54.1% ± 4.6%**	53.9% ± 2.4%	**52.9% ± 2.7%**
MLP	34.8% ± 14.6%	70.4% ± 13.5%	52.4% ± 5.4%	53.9% ± 4.2%	52.6% ± 4.2%
LDA	**41.4% ± 6.8%**	*56.6% ± 3.1%*	49.9% ± 4.0%	49.6% ± 4.0%	49.0% ± 4.2%
KNN	39.5% ± 6.7%	57.1% ± 4.6%	49.2% ± 5.2%	*48.9% ± 3.9%*	*48.3% ± 4.0%*

Table F.34: Multisite classification performance on DF27.

Classifier	Sensitivity	Specificity	AUROC	Accuracy	Balanced Accuracy
LR	45.2% ± 3.3%	57.9% ± 4.4%	53.7% ± 2.6%	52.0% ± 2.5%	51.5% ± 2.4%
SVM	47.2% ± 4.2%	55.3% ± 0.7%	53.7% ± 2.1%	51.5% ± 1.8%	51.2% ± 1.9%
DT	**46.4% ± 8.7%**	54.3% ± 3.5%	51.1% ± 3.9%	50.6% ± 4.4%	50.3% ± 4.7%
RF	*23.1% ± 5.3%*	**76.5% ± 3.6%**	51.8% ± 5.5%	51.8% ± 3.4%	49.8% ± 3.4%
MLP	46.1% ± 4.9%	60.1% ± 8.1%	**55.6% ± 4.2%**	**53.6% ± 3.1%**	**53.1% ± 2.8%**
LDA	45.9% ± 4.0%	*52.6% ± 4.3%*	*49.9% ± 3.1%*	*49.5% ± 2.6%*	*49.2% ± 2.5%*
KNN	40.7% ± 7.7%	60.9% ± 8.6%	50.3% ± 3.8%	51.5% ± 2.8%	50.8% ± 2.7%

F.2.2 Per-site Performances

Table F.35: The average performance across folds on DF6 graph features, evaluated on CALTECH and CMU.

	CALTECH					**CMU**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	66.7±57.7	80.0±34.6	66.7±57.7	77.8±38.5	73.3±46.2	33.3±47.1	**100.0±0.0**	91.7±11.8	65.0±21.2	66.7±23.6
SVM	66.7±57.7	**86.7±23.1**	73.3±46.2	83.3±28.9	76.7±40.4	66.7±47.1	**100.0±0.0**	91.7±11.8	**80.0±28.3**	**83.3±23.6**
DT	**100.0±0.0**	83.3±28.9	91.7±14.4	**88.9±19.2**	**91.7±14.4**	50.0±70.7	75.0±35.4	70.8±29.5	65.0±21.2	62.5±17.7
RF	33.3±57.7	**86.7±23.1**	**93.3±11.5**	61.1±9.6	60.0±17.3	16.7±23.6	75.0±35.4	83.3±23.6	45.0±7.1	45.8±5.9
MLP	33.3±57.7	63.3±32.1	50.0±50.0	55.6±38.5	48.3±44.8	33.3±47.1	**100.0±0.0**	**100.0±0.0**	65.0±21.2	66.7±23.6
LDA	66.7±57.7	70.0±26.5	50.0±50.0	72.2±25.5	68.3±35.5	16.7±23.6	**100.0±0.0**	83.3±23.6	55.0±7.1	58.3±11.8
KNN	0.0±0.0	80.0±34.6	61.7±46.5	50.0±16.7	40.0±17.3	50.0±70.7	75.0±35.4	75.0±35.4	60.0±56.6	62.5±53.0

Table F.36: The average performance across folds on DF6 graph features, evaluated on KKI and LEUVEN_1.

	KKI					**LEUVEN_1**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	64.6±41.0	42.9±23.8	59.4±16.9	48.6±2.8	53.8±9.2	38.3±27.4	68.3±20.7	49.7±30.7	55.1±17.3	53.3±16.5
SVM	64.6±41.0	42.9±17.3	**72.3±14.4**	49.0±12.0	53.8±16.4	26.7±25.3	68.3±20.7	53.9±20.0	48.4±13.8	47.5±13.0
DT	45.8±41.7	44.6±24.4	41.7±13.6	43.4±19.4	45.2±20.4	**53.3±36.1**	**78.3±21.7**	51.9±17.7	**68.0±13.1**	**65.8±11.2**
RF	70.8±21.0	54.2±11.7	66.1±18.5	60.1±10.6	62.5±12.7	16.7±23.6	73.3±43.5	49.2±33.6	45.6±28.5	45.0±29.8
MLP	**70.8±34.4**	49.2±20.6	59.6±28.5	55.6±7.9	60.0±10.8	15.0±22.4	65.0±41.8	43.3±39.7	41.2±28.9	40.0±27.1
LDA	64.6±41.0	54.2±11.7	67.5±24.4	55.9±10.4	59.4±15.6	31.7±32.5	56.7±9.1	45.6±30.5	46.0±18.6	44.2±18.1
KNN	45.8±41.7	**93.8±12.5**	65.4±27.6	**72.2±19.2**	69.8±21.3	33.3±31.2	73.3±25.3	49.7±32.3	54.1±21.9	53.3±20.9

Table F.37: The average performance across folds on DF6 graph features, evaluated on LEUVEN_2 and MAX_MUN.

	LEUVEN_2					**MAX_MUN**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	37.5±47.9	48.8±16.5	42.1±16.5	50.0±5.8	43.1±16.0	58.6±26.2	44.0±26.1	48.8±36.7	55.6±15.9	51.3±21.4
SVM	50.0±57.7	48.8±16.5	57.9±19.6	56.2±13.8	49.4±22.9	52.9±32.7	46.0±36.5	46.2±38.4	49.4±32.7	49.4±34.4
DT	**66.7±47.1**	31.2±34.2	40.0±32.4	47.6±28.4	49.0±32.7	46.2±40.3	**69.0±19.5**	**59.5±14.8**	48.7±21.8	57.6±20.1
RF	62.5±47.9	70.0±24.5	**73.8±20.6**	**68.2±26.8**	**66.2±28.4**	44.5±24.0	59.0±36.8	47.4±28.6	52.1±23.8	51.8±26.5
MLP	0.0±0.0	**77.5±20.6**	46.7±20.5	50.6±16.2	38.8±10.3	55.2±25.9	46.0±27.0	45.2±30.0	53.7±12.2	50.6±20.7
LDA	37.5±47.9	60.0±14.1	42.1±16.5	57.1±10.1	48.8±19.3	48.6±12.2	52.0±37.0	44.4±30.4	54.0±17.3	50.3±22.1
KNN	45.8±41.7	43.8±14.9	59.8±18.9	46.4±12.4	44.8±19.4	**72.6±18.1**	49.0±38.8	51.3±31.7	**63.7±12.1**	**60.8±19.6**

Table F.38: The average performance across folds on DF6 graph features, evaluated on NYU and OHSU.

	NYU					**OHSU**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	48.7±8.4	65.0±8.0	62.5±12.3	58.3±6.1	56.9±5.8	**71.7±29.8**	58.3±37.3	60.8±37.0	66.7±27.0	65.0±31.3
SVM	**57.2±8.1**	64.9±11.3	65.3±9.4	62.1±8.9	61.1±8.1	63.3±21.7	**78.3±21.7**	68.3±12.4	**69.3±11.8**	**70.8±14.4**
DT	36.4±10.9	78.1±6.2	54.9±7.2	59.8±4.6	57.2±3.9	60.0±41.8	56.7±43.5	45.4±16.0	57.7±19.4	58.3±11.8
RF	40.1±19.9	**83.6±6.9**	**66.3±13.9**	**65.2±11.7**	**61.8±13.3**	68.3±32.5	43.3±36.5	40.4±12.6	51.3±9.6	55.8±17.6
MLP	52.8±17.3	69.3±4.9	62.6±11.7	62.5±7.9	61.0±9.0	53.3±36.1	46.7±27.4	45.0±31.9	56.0±13.4	50.0±20.4
LDA	49.0±13.5	64.2±8.2	61.0±13.4	57.9±4.5	56.6±5.3	58.3±37.3	58.3±37.3	62.5±24.7	66.7±20.4	58.3±27.5
KNN	40.4±27.1	75.3±5.7	59.5±13.2	60.5±10.6	57.8±12.7	58.3±37.3	71.7±31.0	**71.7±16.6**	63.7±15.0	65.0±21.8

Table F.39: The average performance across folds on DF6 graph features, evaluated on OLIN and PITT.

	OLIN					**PITT**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	44.2±30.2	58.3±28.9	58.9±30.9	49.3±17.6	51.2±20.6	**67.6±29.2**	73.9±25.4	**74.8±15.4**	**70.6±17.1**	70.7±17.1
SVM	69.2±21.7	64.6±29.2	60.1±10.7	64.6±10.5	66.9±11.8	64.0±38.2	79.1±16.0	70.7±5.3	69.4±10.0	**71.6±12.1**
DT	39.2±28.3	50.0±40.8	48.1±34.8	44.4±34.2	44.6±34.2	23.3±15.5	**86.3±19.2**	49.1±9.7	60.1±13.0	54.8±10.0
RF	45.4±17.5	68.8±23.9	65.8±21.0	55.6±20.8	57.1±19.3	25.4±18.0	79.8±14.0	58.7±25.5	56.2±12.7	52.6±15.2
MLP	69.2±21.7	45.8±8.3	55.5±9.9	57.6±11.9	57.5±11.9	53.3±36.4	82.3±17.7	72.7±24.4	67.2±23.0	67.8±22.5
LDA	37.9±31.2	58.3±28.9	63.3±16.7	45.1±13.7	48.1±16.2	64.7±29.4	57.2±25.8	70.1±18.6	57.5±6.3	61.0±6.0
KNN	**88.8±13.1**	**79.2±25.0**	**82.8±14.5**	**83.3±13.6**	**84.0±14.5**	39.1±23.0	73.2±22.1	61.1±7.1	61.2±11.8	56.2±9.4

Table F.40: The average performance across folds on DF6 graph features, evaluated on SBL and SDSU.

	SBL					**SDSU**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	60.0±43.5	46.2±40.3	48.1±36.4	50.0±32.8	53.1±33.6	16.7±28.9	68.3±16.1	64.2±23.2	48.1±17.0	42.5±18.9
SVM	70.0±44.7	61.9±42.2	**77.0±23.4**	**69.4±19.6**	**66.0±21.0**	25.0±25.0	76.7±25.2	69.2±6.3	57.4±8.5	50.8±1.4
DT	63.3±41.5	39.5±38.7	46.9±17.2	47.2±15.0	51.4±15.9	0.0±0.0	**91.7±14.4**	16.7±19.1	57.4±8.5	45.8±7.2
RF	50.0±47.1	**64.8±41.0**	60.3±26.4	52.5±13.0	57.4±14.9	25.0±25.0	83.3±28.9	39.6±23.7	61.1±9.6	54.2±7.2
MLP	73.3±43.5	29.0±29.8	61.9±28.4	52.5±18.1	51.2±16.6	41.7±38.2	83.3±28.9	68.3±27.0	**68.5±30.6**	**62.5±33.1**
LDA	60.0±43.5	46.2±40.3	51.0±37.9	50.0±32.8	53.1±33.6	25.0±25.0	68.3±16.1	60.0±21.4	51.9±17.0	46.7±19.4
KNN	**83.3±23.6**	48.6±50.1	61.7±26.1	63.3±21.7	**66.0±25.4**	25.0±25.0	**91.7±14.4**	**72.1±18.8**	66.7±16.7	58.3±19.1

Table F.41: The average performance across folds on DF6 graph features, evaluated on STANFORD and TRINITY.

	STANFORD					**TRINITY**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	39.6±31.5	37.5±25.0	49.0±27.7	41.5±18.6	38.5±21.3	41.0±27.0	68.3±20.7	44.8±34.3	59.3±16.4	54.7±19.8
SVM	39.6±31.5	50.0±40.8	35.4±33.6	49.8±21.1	44.8±19.7	37.0±38.0	36.7±13.9	34.2±36.4	40.8±18.8	36.8±18.0
DT	47.9±44.3	31.2±37.5	50.5±12.5	47.5±14.5	39.6±12.5	**60.0±23.5**	35.0±19.0	**53.2±8.2**	42.0±8.7	47.5±12.0
RF	20.8±14.4	37.5±25.0	45.1±25.0	31.0±7.6	29.2±10.2	45.0±28.3	58.3±30.0	50.3±30.9	52.3±21.1	51.7±24.9
MLP	45.8±36.3	43.8±31.5	28.6±30.1	49.6±18.3	44.8±21.3	38.0±37.5	**76.7±34.6**	48.2±29.6	**63.1±10.1**	57.3±5.3
LDA	45.8±41.7	37.5±25.0	52.1±24.7	44.6±23.7	41.7±26.4	37.0±28.6	65.0±22.4	43.5±30.4	55.5±18.9	51.0±21.3
KNN	**60.4±42.7**	25.0±28.9	51.6±23.0	49.6±18.3	42.7±22.1	32.0±21.7	53.3±13.9	39.7±24.7	46.8±10.1	42.7±13.5

Table F.42: The average performance across folds on DF6 graph features, evaluated on UCLA_1 and UCLA_2.

	UCLA_1					**UCLA_2**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	65.0±9.1	57.7±24.4	59.2±9.5	59.1±3.9	**61.3±8.9**	**90.0±14.1**	45.8±29.5	70.8±5.9	65.3±13.7	67.9±21.8
SVM	67.2±10.8	54.9±29.3	55.0±11.0	**59.5±7.9**	61.1±12.1	**90.0±14.1**	**58.3±11.8**	**87.5±17.7**	**70.8±5.9**	**74.2±13.0**
DT	35.3±16.0	**66.8±19.5**	47.1±16.4	48.0±12.9	51.0±17.1	50.0±70.7	37.5±53.0	50.8±22.4	29.2±5.9	43.8±8.8
RF	51.4±19.2	65.0±27.4	63.0±17.2	55.9±10.7	58.2±14.4	80.0±28.3	54.2±29.5	72.1±7.7	58.3±11.8	67.1±0.6
MLP	56.7±9.1	50.9±33.0	55.7±19.4	52.3±9.7	53.8±14.9	**90.0±14.1**	45.8±29.5	75.8±13.0	65.3±13.7	67.9±21.8
LDA	**68.3±12.4**	50.8±31.2	58.2±10.7	58.0±10.2	59.6±13.5	80.0±28.3	45.8±29.5	68.3±2.4	59.7±21.6	62.9±28.9
KNN	52.2±12.2	49.5±28.7	**64.3±9.7**	54.1±7.2	50.9±13.2	70.0±42.4	54.2±29.5	73.3±9.4	52.8±3.9	62.1±6.5

Table F.43: The average performance across folds on DF6 graph features, evaluated on UM_1 and UM_2.

	UM_1					**UM_2**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	44.0±19.9	67.8±12.8	61.0±18.0	59.2±14.6	55.9±14.9	42.7±39.3	86.0±14.2	60.0±38.1	71.4±6.9	64.3±14.1
SVM	52.5±11.4	67.8±15.0	63.2±14.7	61.9±11.1	60.2±10.3	36.0±37.0	82.7±20.5	**78.0±21.8**	66.2±21.4	59.3±24.8
DT	54.8±33.2	**70.4±26.8**	60.8±21.7	**67.7±16.0**	62.6±16.8	38.7±42.5	60.7±29.1	52.7±34.3	53.3±22.8	49.7±24.8
RF	48.0±15.3	69.3±22.7	**72.9±17.8**	61.5±15.0	58.6±14.0	34.7±40.9	56.0±33.9	58.9±35.8	48.0±33.1	45.3±34.5
MLP	**73.3±29.5**	62.8±20.7	71.4±17.5	65.7±19.7	**68.1±20.9**	29.3±40.4	85.3±16.4	63.1±17.6	66.1±17.3	57.3±21.0
LDA	42.3±20.5	67.8±11.4	59.1±17.8	58.2±14.2	55.0±14.5	42.7±39.3	**89.3±9.8**	71.1±21.1	**73.6±6.8**	**66.0±14.9**
KNN	58.0±24.4	60.8±17.5	63.5±15.4	58.8±7.8	59.4±9.8	40.0±43.5	58.0±17.1	54.9±30.8	52.3±25.7	49.0±29.3

Table F.44: The average performance across folds on DF6 graph features, evaluated on USM and YALE.

	USM					**YALE**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	71.0±13.6	49.7±30.7	68.5±11.4	62.1±10.1	60.3±10.8	**81.7±20.7**	76.7±32.5	**86.8±19.5**	**74.2±19.5**	**79.2±17.7**
SVM	73.8±12.6	39.7±10.8	66.4±16.2	62.0±10.1	56.7±8.1	70.0±21.7	52.7±39.2	73.7±18.5	59.1±18.8	61.3±26.4
DT	50.3±16.8	44.3±21.9	44.2±14.9	48.7±8.2	47.3±6.4	40.0±38.4	**86.0±21.9**	78.8±11.3	57.6±18.1	63.0±11.4
RF	58.8±16.0	52.0±20.9	54.1±16.5	56.4±4.8	55.4±7.3	43.3±30.3	82.7±16.7	74.3±25.2	53.4±23.2	63.0±12.0
MLP	67.2±24.6	38.0±13.9	64.2±12.5	55.7±10.1	52.6±9.0	68.3±41.0	72.7±30.0	78.5±25.6	64.5±24.4	70.5±27.0
LDA	**74.3±9.4**	49.7±23.0	**70.8±8.3**	**64.2±10.3**	**62.0±10.4**	71.7±18.3	76.0±25.1	83.8±19.5	67.3±11.5	73.8±14.5
KNN	44.2±11.6	**54.7±20.4**	44.5±19.5	49.0±14.7	49.4±15.3	38.3±31.0	68.0±29.5	56.2±22.0	46.8±10.5	53.2±16.9

F.2.3 Per-age Performances

Table F.45: The average performance across folds on DF6 graph features, evaluated by age group 0-11 and 12-18.

	0-11					**12-18**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	50.9±14.2	51.4±10.2	54.7±16.8	51.2±11.3	51.2±11.7	**58.7±3.0**	65.4±5.4	66.3±4.0	62.2±3.0	62.1±2.7
SVM	**57.2±15.8**	50.8±7.6	56.9±12.2	52.7±9.1	54.0±9.4	57.8±7.2	65.8±6.7	65.5±7.3	62.0±6.1	61.8±6.3
DT	40.6±14.7	**67.6±11.3**	**58.0±5.8**	**56.2±5.2**	**54.1±5.7**	43.3±12.9	64.3±11.0	52.0±4.6	54.0±3.3	53.8±3.6
RF	47.1±13.9	58.8±11.8	55.9±9.6	52.7±8.6	53.0±8.0	42.7±8.5	**71.0±3.5**	63.3±4.2	57.7±4.5	56.8±5.2
MLP	52.2±13.4	50.4±7.0	52.6±13.2	51.4±8.0	51.3±8.9	58.0±14.0	66.8±6.5	**67.5±9.5**	**62.8±8.9**	**62.4±9.5**
LDA	54.3±13.4	51.8±9.8	54.2±15.1	52.9±9.7	53.1±10.2	56.6±5.2	62.8±7.2	65.5±4.7	59.8±4.3	59.7±4.0
KNN	53.0±10.3	55.0±8.7	54.6±8.1	53.7±8.8	54.0±8.2	46.8±8.2	65.5±7.3	60.5±5.6	56.6±4.6	56.2±5.0

Table F.46: The average performance across folds on DF6 graph-based features, evaluated by age group 19-30 and 30+.

	19-30					**30+**				
	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]	SEN [%]	SPE [%]	AUROC [%]	ACC [%]	BACC [%]
LR	51.7±6.5	61.1±6.4	56.7±9.1	57.2±5.4	56.4±6.0	50.3±19.6	54.8±29.6	54.4±28.6	51.8±18.1	52.6±20.6
SVM	**53.4±14.6**	61.0±9.8	58.9±10.7	58.4±7.4	57.2±8.9	55.9±29.8	72.7±18.6	**69.5±17.7**	**63.8±21.1**	**64.3±18.6**
DT	43.2±20.2	62.5±14.5	46.7±9.6	54.8±2.8	52.9±5.8	54.1±14.0	**72.8±27.1**	63.4±18.2	58.4±12.8	63.4±16.8
RF	42.9±3.9	**75.6±10.3**	**62.0±4.2**	**61.7±4.6**	**59.2±4.4**	44.3±18.5	65.2±10.7	58.4±8.4	54.1±12.8	54.7±9.8
MLP	46.3±14.6	61.8±4.4	54.4±10.0	55.7±6.7	54.0±8.2	51.5±27.8	64.0±24.9	58.7±27.3	57.8±24.9	57.7±25.4
LDA	47.5±10.4	61.2±9.7	56.2±9.4	55.7±8.7	54.4±9.6	43.7±18.3	57.8±27.4	52.7±30.6	50.6±17.7	50.8±20.6
KNN	42.2±15.9	69.0±8.5	54.2±11.5	58.5±8.2	55.6±9.2	**59.6±15.2**	67.8±19.0	63.0±7.2	**63.8±9.0**	63.7±11.8

F.2.4 LOGO CV on Graph features

Table F.47: Sensitivity obtained by each classifier for each test site during LOGO CV on graph features.

	Sensitivity [%]							**N**	
	LR	SVM	DT	RF	MLP	LDA	KNN	ASD	TC
CALTECH	**60.0**	**60.0**	40.0	0.0	**60.0**	**60.0**	20.0	5	10
CMU	**50.0**	**50.0**	16.7	16.7	33.3	33.3	16.7	6	5
KKI	75.0	66.7	58.3	66.7	**83.3**	**83.3**	33.3	12	21
LEUVEN_1	28.6	21.4	**85.7**	21.4	14.3	21.4	21.4	14	14
LEUVEN_2	16.7	41.7	50.0	25.0	33.3	16.7	**75.0**	12	16
MAXMUN	52.6	47.4	**73.7**	47.4	42.1	42.1	57.9	19	27
NYU	45.9	51.4	25.7	45.9	41.9	**54.1**	39.2	74	98
OHSU	75.0	58.3	58.3	75.0	**91.7**	66.7	75.0	12	13
OLIN	64.3	50.0	**71.4**	64.3	64.3	64.3	**71.4**	14	14
PITT	**62.5**	37.5	54.2	20.8	58.3	58.3	45.8	24	26
SBL	50.0	41.7	50.0	50.0	50.0	**58.3**	41.7	12	14
SDSU	**50.0**	37.5	12.5	25.0	37.5	37.5	**50.0**	8	19
STANFORD	75.0	**83.3**	75.0	41.7	75.0	**83.3**	58.3	12	13
TRINITY	36.8	63.2	**84.2**	63.2	36.8	36.8	42.1	19	25
UCLA_1	70.3	67.6	62.2	37.8	62.2	**73.0**	48.6	37	27
UCLA_2	63.6	63.6	45.5	63.6	72.7	63.6	**81.8**	11	10
UM_1	44.1	47.1	50.0	44.1	**70.6**	41.2	58.8	34	52
UM_2	69.2	76.9	**92.3**	69.2	38.5	61.5	53.8	13	21
USM	67.4	51.2	53.5	44.2	**83.7**	74.4	34.9	43	24
YALE	**77.3**	63.6	50.0	31.8	**77.3**	72.7	40.9	22	19
Mean±std	56.7±16.2	54.0±14.3	55.5±21.0	42.7±19.9	56.3±20.5	55.1±18.8	48.3±18.1		

Table F.48: Specificity obtained by each classifier for each test site during LOGO CV on graph features.

	Specificity [%]							**N**	
	LR	SVM	DT	RF	MLP	LDA	KNN	ASD	TC
CALTECH	70.0	**80.0**	**80.0**	70.0	50.0	60.0	50.0	5	10
CMU	80.0	60.0	80.0	**100.0**	60.0	80.0	60.0	6	5
KKI	38.1	38.1	61.9	38.1	33.3	33.3	**66.7**	12	21
LEUVEN_1	**100.0**	57.1	42.9	78.6	85.7	**100.0**	71.4	14	14
LEUVEN_2	68.8	75.0	50.0	**81.2**	56.2	75.0	43.8	12	16
MAXMUN	51.9	44.4	55.6	**63.0**	59.3	55.6	**63.0**	19	27
NYU	73.5	79.6	**83.7**	78.6	74.5	69.4	68.4	74	98
OHSU	69.2	53.8	61.5	30.8	**76.9**	**76.9**	69.2	12	13
OLIN	57.1	**64.3**	28.6	**64.3**	**64.3**	**64.3**	57.1	14	14
PITT	46.2	57.7	69.2	**92.3**	46.2	42.3	65.4	24	26
SBL	42.9	57.1	50.0	**78.6**	50.0	50.0	57.1	12	14
SDSU	84.2	68.4	**94.7**	78.9	73.7	84.2	78.9	8	19
STANFORD	38.5	46.2	38.5	**69.2**	53.8	38.5	53.8	12	13
TRINITY	64.0	60.0	12.0	52.0	48.0	**68.0**	48.0	19	25
UCLA_1	63.0	63.0	59.3	**77.8**	63.0	66.7	70.4	37	27
UCLA_2	50.0	60.0	60.0	**80.0**	50.0	50.0	50.0	11	10
UM_1	65.4	69.2	55.8	**71.2**	51.9	65.4	57.7	34	52
UM_2	81.0	85.7	52.4	71.4	**90.5**	81.0	66.7	13	21
USM	50.0	62.5	62.5	**66.7**	16.7	45.8	50.0	43	24
YALE	73.7	68.4	36.8	**78.9**	47.4	73.7	63.2	22	19
Mean±std	63.4±16.1	62.5±11.8	56.8±19.2	71.1±15.8	57.6±16.9	64.0±16.7	60.5±9.1		

Table F.49: Auroc obtained by each classifier for each test site during LOGO CV on graph features.

	AUROC [%]							**N**	
	LR	SVM	DT	RF	MLP	LDA	KNN	ASD	TC
CALTECH	58.0	62.0	**63.0**	57.0	58.0	50.0	27.0	5	10
CMU	63.3	60.0	66.7	51.7	**73.3**	70.0	55.0	6	5
KKI	63.9	66.3	**66.9**	53.6	58.3	65.1	54.4	12	21
LEUVEN_1	58.7	37.8	**67.9**	61.7	52.0	56.1	48.0	14	14
LEUVEN_2	51.0	**67.2**	52.6	50.5	55.2	52.6	58.1	12	16
MAXMUN	46.2	45.7	**68.7**	56.3	52.0	46.0	61.0	19	27
NYU	66.6	**68.8**	60.6	65.3	61.2	67.6	56.8	74	98
OHSU	82.1	50.6	64.4	48.4	**86.5**	78.2	73.1	12	13
OLIN	67.9	63.5	49.0	**70.7**	66.3	65.3	66.3	14	14
PITT	61.9	52.6	56.4	**69.1**	54.6	58.3	56.1	24	26
SBL	51.2	55.4	**62.5**	**62.5**	47.6	54.2	48.2	12	14
SDSU	67.1	67.1	56.9	66.1	**70.4**	69.7	67.1	8	19
STANFORD	**67.3**	59.6	57.1	59.0	64.7	64.1	56.4	12	13
TRINITY	53.9	**57.5**	49.4	49.6	39.4	52.4	39.5	19	25
UCLA_1	67.9	**69.3**	63.1	64.1	67.4	67.5	61.3	37	27
UCLA_2	65.5	69.1	53.2	**77.7**	66.4	55.5	70.9	11	10
UM_1	57.5	59.3	56.5	64.5	**64.6**	55.6	63.8	34	52
UM_2	82.1	**86.1**	73.3	72.7	75.8	84.2	63.9	13	21
USM	**70.3**	58.7	57.3	53.2	54.2	69.2	40.3	43	24
YALE	**75.6**	75.1	42.9	66.5	68.7	71.5	51.6	22	19
Mean±std	63.9±9.3	61.6±10.3	59.4±7.5	61.0±8.1	61.8±10.5	62.7±9.7	55.9±11.0		

Table F.50: Accuracy obtained by each classifier for each test site during LOGO CV on graph features.

	Accuracy [%]							**N**	
	LR	SVM	DT	RF	MLP	LDA	KNN	ASD	TC
CALTECH	66.7	**73.3**	66.7	46.7	53.3	60.0	40.0	5	10
CMU	**63.6**	54.5	45.5	54.5	45.5	54.5	36.4	6	5
KKI	51.5	48.5	**60.6**	48.5	51.5	51.5	54.5	12	21
LEUVEN_1	**64.3**	39.3	**64.3**	50.0	50.0	60.7	46.4	14	14
LEUVEN_2	46.4	**60.7**	50.0	57.1	46.4	50.0	57.1	12	16
MAXMUN	52.2	45.7	**63.0**	56.5	52.2	50.0	60.9	19	27
NYU	61.6	**67.4**	58.7	64.5	60.5	62.8	55.8	74	98
OHSU	72.0	56.0	60.0	52.0	**84.0**	72.0	72.0	12	13
OLIN	60.7	57.1	50.0	**64.3**	**64.3**	**64.3**	**64.3**	14	14
PITT	54.0	48.0	**62.0**	58.0	52.0	50.0	56.0	24	26
SBL	46.2	50.0	50.0	**65.4**	50.0	53.8	50.0	12	14
SDSU	**74.1**	59.3	70.4	63.0	63.0	70.4	70.4	8	19
STANFORD	56.0	**64.0**	56.0	56.0	**64.0**	60.0	56.0	12	13
TRINITY	52.3	**61.4**	43.2	56.8	43.2	54.5	45.5	19	25
UCLA_1	67.2	65.6	60.9	54.7	62.5	**70.3**	57.8	37	27
UCLA_2	57.1	61.9	52.4	**71.4**	61.9	57.1	66.7	11	10
UM_1	57.0	**60.5**	53.5	**60.5**	59.3	55.8	58.1	34	52
UM_2	76.5	**82.4**	67.6	70.6	70.6	73.5	61.8	13	21
USM	61.2	55.2	56.7	52.2	59.7	**64.2**	40.3	43	24
YALE	**75.6**	65.9	43.9	53.7	63.4	73.2	51.2	22	19
Mean±std	60.8±9.0	58.8±9.7	56.8±7.8	57.8±6.7	57.9±9.4	60.4±7.9	55.1±9.6		

G NASDA Prototype Documentation

G.1 Overview

The Neuroimaging Autism Signal Detection Application (NASDA) prototype is a modular graphical user interface (GUI) designed to assist in exploring pattern recognition of neuroimaging-derived features. It is designed to enable non-specialist users and research scientists to rapidly adjust, run, and inspect multiple classification scenarios without requiring manual scripting.
The prototype integrates subgroup selection, classifier choice, feature selection, and direct performance monitoring into a single visual workspace. All operations are implemented in Python using `Tkinter` [36] for the GUI framework, `scikit-learn` [29] for classification routines, and `nilearn` [22] for neuroimaging data handling.

G.2 Layout and Components

The NASDA prototype GUI is organized to maximize transparency, reproducibility, and flexible experiment configuration. Its modular design supports immediate visual feedback and reliable user control. Each panel with its functions is described below.

- **Toolbar:** A persistent strip at the top of the window providing global project controls.
  - `Open:` Load a new dataset in a new window.
  - `Save:` A placeholder button that can be linked to saving current configurations to a dedicated project file for later use.
  - `Run:` Execute the entire classification and evaluation workflow, respecting the current settings and filters.
  - `Settings:` Open a dedicated window to adjust advanced runtime parameters. The settings panel currently supports a user-editable text entry, that specifies and stores the custom file path for the graph-based feature file.
- **Subjects Panel:** Enable dynamic dataset subsetting based on demographic metadata.
  - `Sex Filters:` Radio buttons for *All*, *Female*, or *Male*. Changing this filter triggers an update of the analysed dataset.
  - `Age Group Filters:` Radio buttons for *All*, *0–11*, *12–18*, *19–30*, and *30+*. Used in tandem with the sex filter for refined subgroup analyses.
- **Classifier Panel:** Specify the applied classification model.
  - Available options: `SVM`, `Logistic Regression`, `Random Forest`, `Decision Tree`, `Multilayer Perceptron`, `Linear Discriminant Analysis`, and `k-Nearest Neighbour`.
  - Selection is mutually exclusive, managed via radio buttons.
- **Features Panel:** Configure both the feature space and the feature selection procedure.
  - `Feature Selection Methods:` Choices include `None`, `Cluster`, `Lars Lasso`, `HSIC Lasso`, and `Backwards Sequential Feature Selection (SFS)`.
  - `Feature Type:` Toggle between *Graph-based signal features* or *Pearson correlation matrices.* Switching this option dynamically reloads the input data representation to match the selected modality.
- **Overview Visualisation:** A dedicated graphical panel to display a stylised brain schematic with overlaid text summarizing:
  - Active demographic filters (sex and age group),
  - Selected classifier,
  - Selected feature method,

  - Chosen feature type and data source.

  Two utility buttons are included in this visualisation plane: one exports the panel as a PNG image, and the other enlarges the view for closer inspection.

- **Tabbed Console:** A multi-tab control for interactive workflow management, inspection, and debugging.
  - `Command Tab:` A live Python interpreter shell embedded directly in the GUI, allowing direct data inspection, and method calls. All script output is shown in real-time.
  - `Error Log:` Collects Python errors, exceptions, and tracebacks for efficient troubleshooting.
  - `Data Fitting Tab:` Displays current status of the cross-validation procedure, including fold progression and split summaries.
  - `Performance Tab:` Presents detailed performance metrics such as accuracy, sensitivity, specificity, AUROC, confusion matrices, and per-fold statistics.
  - `Help Function:` A built-in `help()` command prints an indexed summary of all recognized commands, classifier descriptions, and usage instructions.

Available Runtime Functions:
In addition to graphical controls, the console supports direct execution of defined helper functions:

- `run()`: Executes the full pipeline using the current filters, classifier, and feature settings.
- `settings()`: Opens a dedicated window to adjust advanced runtime parameters.
- `runanalysis(stats)`: Simulates an analysis for demonstration and debugging purposes.
- `log('message')`: Prints a custom user message to the console area.
- `export_overview_to_png()`: Saves the current overview visualisation as an external image file.
- `help()`: Outputs an updated list of available commands, classifiers, and recommended usage.

This comprehensive configuration, including persistent user-editable file paths and context-aware dataset reloading, reinforces NASDA's aim of providing a reliable yet flexible neuroimaging pipeline accessible to both novice users and research engineers.

G.3 Typical Workflow

1. **Load or configure dataset:** By default, the prototype loads the ABIDE I dataset; future versions could support user-uploaded CSVs or integrated `nilearn` data fetchers.
2. **Define subgroup(s):** Select desired demographic filters (sex, age group). This dynamically re-filters the dataset.
3. **Choose classifier and features:** Pick one classifier and desired feature selection method(s). The GUI automatically updates internal parameters.
4. **Run analysis:** Click the *Run* button or use the embedded command line to execute the pipeline. The backend performs stratified cross-validation, applies feature selection, if specified, fits the model, and prints the performance metrics.
5. **Inspect results:** Performance metrics (accuracy, recall, AUROC, confusion matrix) are streamed live to the console. Logs can be saved for reporting. The command can give further insights with standard python commands, such as `print(y.size)`.

G.4 Limitations

The prototype is not a clinical diagnostic tool. It assumes that users understand cross-validation principles and feature selection biases. It should be employed strictly for hypothesis exploration and preliminary testing before formal statistical validation.

G.5 System Requirements

- Python 3.10+ with `Tkinter`, `scikit-learn`, `nilearn`, `Pandas`, and `Pillow`.
- Windows desktop environment with standard display resolution.
- Sufficient RAM for handling full correlation matrices.

H Python Codes

All codes used in this research can be found on: `https://github.com/cxwchen/AutismDetection`.