# An integrated approach to characterize Autism Spectrum Disorder

by

Ferran Gonzalez Hernandez

CoMPLEX Summer Project Thesis

Supervisors: Ilias Tachtsidis & Paul Burgess

University College London

August 2018

# *Abstract*

Autism spectrum disorder (ASD) is a neurodevelopmental condition with a complex pattern of deficits at multiple levels. Common studies have traditionally focused on detecting group differences of specific measures between typically developed (TD) and ASD participants. However, most findings report a high degree of heterogeneity among affected individuals, presenting an uneven pattern of deficits at the behavioural, genetic and neural levels across the whole spectrum. This study aimed to investigate the discriminant capacity of multivariate approaches to identify ASD at the subject level, considering different aspects of behavioural and neural deficits. Moreover, it was studied whether the integration of behavioural and neuroimaging data provided complementary information across domains and improved the identification of ASD. A number of behavioural and neuroimaging features were extracted from 26 TD and 26 participants with high-functioning ASD, and they were used as an input for a machine learning classifier. The results showed significant discriminance between TD and ASD subjects during leave-one out classification approaches and exhibited the advantages of this methodology against traditional univariate techniques. Four behavioural and seven neuroimaging discriminant features were detected by the algorithm, and their role as functional biomarkers was discussed. Finally, the integration of behavioural and neuroimaging features did not improve the classification performance, suggesting that features extracted in both domains provided similar information.

# Contents

**Abstract** **i**

**1 Introduction** **1**
1.1 Cognitive neuroscience in ASD . . . 2
1.1.1 Functional Near-Infrared Spectroscopy . . . 3
1.2 Previous work and new challenges . . . 4
1.2.1 Machine learning for single subject prediction . . . 6
1.2.2 Multimodal approaches . . . 8
1.3 Context and experimental aims . . . 8

**2 Methods** **10**
2.1 Experimental Design . . . 11
2.2 fNIRS data acquisition and pre-processing . . . 12
2.3 Feature extraction . . . 12
2.3.1 Behavioural features . . . 13
2.3.2 fNIRS features . . . 14
2.4 Classification and feature selection . . . 18
2.4.1 LogitBoost algorithm . . . 19
2.4.2 Feature selection . . . 22
2.5 Performance evaluation . . . 25

**3 Results** **27**
3.1 Data visualization and summary statistics . . . 27
3.1.1 Behavioural features . . . 27
3.1.2 fNIRS features . . . 32
3.2 Classification . . . 34
3.2.1 Behavioural analysis . . . 34
3.2.2 fNIRS analysis . . . 35
3.2.3 Integrated analysis . . . 37

**4 Discussion** **39**
4.1 Behavioural discriminance . . . 40
4.2 fNIRS discriminance . . . 41
4.3 Integrated classification . . . 43

**5 Conclusions** **45**

**Bibliography** **46**

# Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition affecting around 1% of the population worldwide [1]. Autism is characterized by a triad of deficits, including impairments in social interaction, communication difficulties, along with stereotyped interests and behaviours [2].

Over the last decades, significant research has been conducted in the field of ASD, which led to the identification of certain characteristic patterns of this condition across different domains. For example, ASD is considered to be a highly heritable condition, which has triggered a number of studies in the field of genetics. Several genes have been found to be associated with ASD, which together account for 10-20% of autistic cases [3]. Moreover, neuroanatomical studies have focused on finding structural abnormalities in autistic brains, and differences in brain volume and white matter connectivity have already been reported [4, 5]. Finally, one of the most promising areas is the field of cognitive neuroscience, which has given insights into the cognitive deficits of ASD subjects together with the associated neural substrates. For instance, behavioural differences have become particularly evident in ASD subjects when performing executive functions (EF; e.g., working memory, planning, cognitive flexibility, inhibition), and multiple functional neuroimaging studies have reported atypical brain activity in ASD during EF tasks [6, 7]. However, the variety of cognitive tests and experimental protocols have often reported conflicting results about the atypical relation brain-behaviour in ASD.

Although significant progress has been made and specific findings start to converge, most studies reveal a high degree of heterogeneity among individuals with ASD across and within domains. This fact results in a group of atypical subjects identified in each study, but it is still far from an adequate characterization of the whole spectrum. Nowadays, the diagnosis of ASD is entirely based on behavioural observations and clinical interviews which is easy and fast to perform in the clinical setting but has often resulted in misdiagnosed subjects and similar treatments applied to a highly heterogeneous group [8]. One of the main limitations behind the development of functional biomarkers for ASD diagnosis has probably been the analysis of single features within one domain. A large number of studies have identified genetic, cognitive and brain features that, in isolation, reported statistically significant differences at the group level, but provided little discriminant capacity when trying to classify subjects at the individual level.

Given the neurobiological heterogenety of ASD, an adequate identification at the subject level would probably require the analysis of more than one feature at a time, looking at variability between and within subjects. Therefore, the diagnosis of ASD is likely to require an integrative platform, focusing on characteristic patterns at the subject level, and considering the complex pattern of deficits across and within domains [2, 7, 8].

For this study, different behavioural and neuroimaging features were extracted and integrated into a single platform to study the predictive power of multifactorial data as a diagnostic tool. In this section, an introduction into the cognitive and neuroimaging topics behind this study will be first presented. Then, the promises and pitfalls of multivariate pattern classification approaches for single subject prediction are disused, and the context and experimental aims of this specific study are finally exposed.

## 1.1 Cognitive neuroscience in ASD

Over the last decades, cognitive neuroscience has played a crucial role towards understanding the relation between human behaviour and those processes that involve cognition [9]. Classically, and particularly in the field of cognitive neuropsychology, organizational models of the cognitive system were built by assessing the behavioural performance of participants during cognitive tasks [9]. More recently, cognitive neuroscience has greatly benefited from functional neuroimaging techniques (e.g., functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG), positron-emission tomography (PET)) which have enabled researchers to map the relation between specific mental functions or cognitive processes with the associated regional brain activity. Due to the number of cognitive impairments found across the whole autism spectrum, these approaches have greatly enhanced our understanding of this condition.

Some of the main cognitive deficits found in ASD include difficulties in: planning, reasoning, attention, working memory, ability to remember complex visual information and others [7]. Particularly, a large number of studies have reported evidence of deficits in ASD during executive activities, not only at the behavioural level but also detecting atypical brain activation patterns together with abnormal interaction between regions when participants performed EF tasks [1]. EFs include a broad range of high-order cognitive processes for organizing and controlling self-behaviour, such as planning, response inhibition, monitoring or multitasking [6], and which play a crucial role in social cognition, moral behaviour and communication [10].

As a result, a variety of cognitive tests and neuroimaging studies on executive functions have been developed [8, 10]. One common finding has been that ASD participants tend to show atypical patterns in only a restricted set of EFs [6]. In addition, different areas in the brain have been associated with executive activities. One specific region that has been recognised to be particularly involved during EF tasks is the prefrontal cortex (PFC). Evidence from multiple studies have reported atypical activation patterns along the PFC of ASD participants during EF tasks [1, 6, 10]. However, none EF-test has yet been found to exhibit deficits in all cases, and instead, an uneven and heterogeneous pattern across and within ASD subjects has often been observed [7]. Moreover, different regions along the PFC support specific aspects of EF, and different regional activations are found depending on the specific task [6]. This variety of tasks together with the high heterogeneity of deficits in ASD has often limit the extraction of general conclusions for the whole spectrum.

Most studies in the field of functional neuroimaging aim to relate specific mental tasks and behaviours with the activity located in particular regions of the brain. However, one of the main limitations of conventional neuroimaging techniques are the experimental set-ups in which the studies can take place, and as a consequence, the behaviours and cognitive tasks that can be studied. The fact that techniques such as fMRI, PET or MEG impose significant physical constrains on the environmental conditions of the experiments (e.g. fMRI scanner) limits the range of situations in which subjects can be studied. For example, the performance of complex tasks that better mimic everyday life conditions - such as the interaction between subjects or multitasking - are hard to replicate in typical neuroimaging laboratories [11].

On the other hand, being able to explore the difficulties that ASD subjects experience when performing EFs in everyday life situations while their neural signal is recorded would open a new range of possibilities. One of the main advantages would be to better investigate the neural underpinnings of EF deficits by means of experiments that involve self-initiated behaviours in naturalistic situations. For this reason, functional near-infrared spectroscopy (fNIRS) is likely to be a particularly suited technique for the study of ASD, as it allows to study cognitive tasks in ecologically-valid and open-ended situations [12].

### 1.1.1 Functional Near-Infrared Spectroscopy

fNIRS is a relatively new neuroimaging technique, which usage has rapidly increased over the last decades, particularly in the field of cognitive neuroscience [11]. Some of the main reasons for this fact are that fNIRS devices are non-invasive, cheap, portable and highly

robust against body movements, which enables the performance of experiments in more naturalistic environments. fNIRS systems measure the hemodynamic response associated to brain activity. In particular, concentration changes of oxygenated ($HbO_2$) and deoxyganated (HHb) haemoglobin are recorded in real time [11]. These measurements are performed by placing a number of light emitters and photodetectors over the scalp and sending light through the biological tissue in the near-infrared range (700-1000 nm). Due to the relative transparency of biological tissue, most variation in light intensity detected by the photodetectors is associated to changes in haemoglobin concentration. Specifically, the emitters send light at two wavelengths ($\lambda_1$, $\lambda_2$) to account for variations in both compounds, $HbO_2$ and HHb. Differences between emitted and detected light intensity are converted to optical density by means of the modified Beer Lambert Law (MBLL) [13]. For a source-detector separation of $d$ and extinction coefficients (at the specific wavelengths) of $\epsilon_{HbO_2\lambda}$ and $\epsilon_{HHb\lambda}$, concentration changes of $HbO_2$ and HHb can be computed with Equation 1.1 [13]:

$$\begin{bmatrix} \Delta[HHb] \\ \Delta[HbO_2] \end{bmatrix} = d^{-1} \begin{bmatrix} \epsilon_{HHb\lambda_1} \epsilon_{HbO_2\lambda_1} \\ \epsilon_{HHb\lambda_2} \epsilon_{HbO_2\lambda_2} \end{bmatrix}^{-1} \begin{bmatrix} \Delta OD(\lambda_1)/DPF(\lambda_1) \\ \Delta OD(\lambda_2)/DPF(\lambda_2) \end{bmatrix} \tag{1.1}$$

Where $OD(\lambda)$ are the changes in optical density for the emitted wavelengths and $DPF(\lambda)$ represents the differential pathlength factor, which varies depending on a number of factors such as the age of the subject, the emitted wavelength or the type of tissue [14]. fNIRS signals are measured from the brain volume located at half of the source detector distance and at a depth of half of the source detector distance. This measurement point is called channel. The number and location of channels used in a particular study depends on the composition of emitters and detectors that the particular device is equipped with [11]. More details about the fNIRS technology and suitability for this study were explained in a previous study presented in the report of the $3^{rd}$ mini-project [15].

## 1.2 Previous work and new challenges

Traditional studies of brain disorders have generally focused on detecting abnormal patterns of certain features in a group of patients in comparison to a healthy cohort [16]. These studies are normally based on univariate analysis, in which a single feature from one domain (either behavioural, genetic, cognitive or neurological) is used at a time. The common methodology consists on averaging parameter values in each of the groups and performing statistical tests to elicit significant differences between patients

and controls by means of p-values [8]. As a result, a number of studies have been published, reporting characteristic patterns in different dimensions of brain disorders.

Despite the undeniable success of this approach in our understanding of multiple brain pathologies, it provides little information when aiming to detect functional biomarkers for clinical diagnosis [8]. The main explanation behind this occurrence is that the identification of specific conditions at the subject level is in fact a different research question, that requires other methodologies than comparing group means. When assessing mean differences between two populations, it is often the case that even if two distributions present significantly different group means, they also have a high degree of overlapping in terms of their range of values. As an example, in Figure 1.1a, the scatter plot of two groups (100 observations each) is illustrated. It is observed how the mean of each group is statistically different from the other, exhibiting a p-value of 0.001. However, if one tries to establish a threshold to classify subjects, only a 60% of the samples can be discriminated. On the other hand, Figure 1.1b shows how the opposite case can occur. Here, if the mean between the two groups is computed, this would range in almost the same value, and the p-value obtained in the t-test would be 0.86. Nonetheless, if two lines are used as a threshold, 94.5% accuracy in classifying the subjects between two categories is obtained, showing the clear outperformance of this methodology in the scenario of single subject prediction against conventional t-test [16].

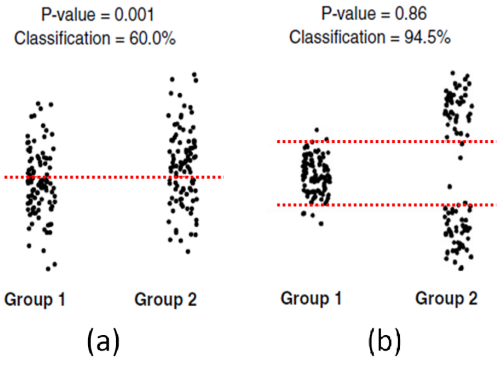



FIGURE 1.1: Illustration of the relative meaning of significant p-values for classification purposes. Group differences are analysed by means of p-values from two-sample t-tests and compared against the classification performance obtained by a simple line threshold (dotted red line). Figure adapted from [16].

For this reason, contemporary techniques that do not average parameter values between groups and, at the same time, can consider multiple features from one modality, become particularly suited for diagnostic purposes, opening the door to single subject-prediction in an automated fashion. These methodologies are normally referred to as multivariate pattern classification or machine learning approaches, and over the last few years, a large number of studies have started to apply these techniques for developing diagnostic tools of complex brain disorders [16]. Moreover, a minor number of studies have started to use machine learning approaches to integrate multiple features from different modalities, combining behavioural, genetic or neuroimaging data [17].

On the other hand, the application of multivariate machine learning techniques also comes at a cost. First, these methodologies have been extensively criticized due to difficulties on interpreting the processes by which an algorithm determines whether a subject has a particular disease/condition. Most of these algorithms were initially developed for technological and industrial purposes, particularly in the field of artificial intelligence, where the functionality of these methods was clearly preferable than their interpretablity. For this reason, the adaptation of these algorithms into the clinical setting would require to provide insights into the bases of their computations/decisions. Moreover, these tools are still relatively new in the field of brain disorders and, in some studies, the inadequate implementation of machine learning algorithms has already led to the extraction of wrong conclusions from the data [16]. In the following sub-section, previous studies using these techniques for brain disorders will be briefly reviewed, and the main pitfalls and challenges will be discussed.

### 1.2.1 Machine learning for single subject prediction

Despite the great capacity of current multivariate approaches to find patterns in highly dimensional datasets, classification at the subject level also represents a much more challenging task than reporting group differences, as it provides information of each individual in the cohort [16]. However, in many occasions these approaches provide highly important information for clinical purposes. One of the main advantages of multivariate approaches is their ability to work and detect interrelated patterns between different features rather than independent points in a single dimension. This fact, opens up the possibility of detecting complex patterns hidden in high dimensional datasets [18]. In the field of neuroimaging, a large number of studies have already reported promising results for the diagnosis of several brain disorders [16], some of them eliciting very high classification accuracies between patients and controls. However, machine learning is still a relatively new domain in neuroimaging and pitfalls are frequent, and the interpretability of results into the context of each particular study and methodology implemented are crucial when extracting general conclusions.

Perhaps one of the areas where machine learning has been applied more often in neuroimaging are MRI studies (with more than 500 papers on single-subject prediction [16]), where highly dimensional datasets are generated and advanced techniques are often required to detect relations between variables. In particular, a large number of studies have focused on analysing the application of multivariate approaches to classify subjects with mild cognitive impairment (MCI), schizophrenia and major depressive disorders (MDD), with many of these studies reporting accuracies above 80 % [16]. In a recent review by Arbabshirani 2017 [16], results from these studies were analysed together with

reports from other brain disorders, including a few papers on attention deficit hyperactivity disorder (ADHD) and ASD. However, one of the main patterns detected in this review was that accuracies often decreased with sample size, questioning the generalization capability of findings from many studies. Moreover, inconsistent methodologies and approaches were found across the literature, and some limitations of machine learning in the neuroimaging field were also identified.

The main limitation of most brain disorder studies lies on the sample size. Most machine learning algorithms have a large number of parameters to optimize in order to build predictive models. However, their ability to optimize those parameters, and therefore to detect general patterns in the population is directly affected by the number of observations that are used as an input for the algorithm. Moreover, the smaller the sample size, the more likely to overfit the data. The word "overfitting" is often used in the machine learning community and it refers to those situations in which predictive models describe noise in the data rather than characteristic patterns. If overfitting occurs, very high accuracies are observed in the data that the algorithm uses for training, but then, poor classification can be achieved on unseen data. Due to the relatively small sample size in neuroimaging studies (compared to other fields where machine learning is applied), overfitting is frequent and conclusions from models with small generalization capability are often extracted [16]. However, even in those scenarios, overfitting can be prevented. In order to avoid this artifact and to extract general conclusions from the data, a number of considerations must be taken into account, specially when dealing with small datasets.

For example, a common methodology is to select those features that are more informative for classification purposes (feature selection). In machine learning studies, feature selection is normally performed in a set of observations (training set) to detect general patterns, and then the performance of the algorithm using this subset of optimal features is evaluated on an unseen group of observations (test set). However, it is not uncommon to see how in some studies dealing with small sample sizes feature selection is performed on the whole dataset [19], given that when splitting the samples into two groups (train and test set), the number of observations is further reduced. One of the main problems of this procedure is that the algorithm could detect a specific dimensional space in which the data can be discriminated with high accuracy but without detecting any generalistic pattern. In those scenarios, the algorithm detects a discriminant dimensional space, but this is due to a random effect rather than a population pattern, which results in a very low generalization capability. Therefore, it is essential that this process is just applied on the training group and later used on a validation set.

Finally, another important fact is that when dealing with small datasets, overall accuracies need to be compared with the random chance of getting that result, to provide in this way a confidence interval for a certain classification accuracy [16]. This is specially important in the field of machine learning classifiers dealing with small datasets. First, because some algorithms have random components within their computations and the confidence intervals between different classifiers might differ. But most importantly, because random chance becomes much greater in small datasets. For example, achieving a classification accuracy of 80% (in a binary classification problem) with 10 observations would not be significantly different from getting 50%, as the 95 % binomial confidence interval would range from 44.4 to 97.5% [16]. Therefore, it is essential to also provide the null distribution of each dataset for each algorithm used.

### 1.2.2 Multimodal approaches

One of the main advantages of applying machine learning platforms on heterogeneous brain disorders such as ASD is that it enables researchers to integrate data from more than one modality and consider in this way the complex pattern of deficits that is often observed across domains. However, in the field of brain disorders, this approach has just been observed in a very small number of papers [17, 18]. Despite the fact that some studies have already speculated about the advantage of integrating multimodal data for the study of ASD [2, 8], to our knowledge, it has never yet been applied. One of the main motivations behind this approach is that whereas one modality might just provide a limited view into one aspect of brain function, behaviour or genetics, multimodal approaches would open-up the possibility to detect hidden patterns across domains.

## 1.3 Context and experimental aims

The study presented in this report is focused on the analysis of experimental data previously collected as part of a collaboration project on autism spectrum disorder between the UCL Institute of Cognitive Neuroscience and the UCL Department of Medical Physics and Biomedical Engineering. Data was collected from a group of typically developed (TD) and ASD subjects, and participants were asked to perform a number of tasks while their neural signal was recorded using an fNIRS device.

Previous analyses on this dataset involved the application of multivariate approaches to discriminate between TD and ASD subjects based on characteristic features of their haemodynamic signal [15]. Those analyses consisted on extracting one feature from

each of the 16 channels of the fNIRS device to use them as an input for a classification algorithm. Different types of features were used to perform separate classification analyses.

In the present study, the capacity to discriminate between TD and ASD subjects at the individual level was investigated using behavioural and fNIRS features as an input for a machine learning classifier. Two feature sets were generated, each one containing 11 behavioural and 38 fNIRS features. Visualization and classification techniques were then applied to each set separately and for an integrated analysis. To our knowledge, this is the first time that different behavioural and fNIRS features are integrated into a single multimodal classification platform, and it enables to study whether this combination provides additional information of ASD deficits across domains.

The main objectives of this study can be summarized: (1) analyse whether multivariate approaches improve the characterization of ASD when different aspects of each modality are simultaneously considered, (2) identify informative features on the basis of behavioural and fNIRS data, (3) study the capacity of certain behavioural and fNIRS features as functional biomarkers to discriminate between TD and ASD subjects, and (4) investigate whether the integration of behavioural and neuroimaging data provides complementary information for a better characterization of ASD at the subject level.

# Methods

Behavioural and fNRIS neuroimaging data were obtained from a group of individuals with ASD and TD subjects. During the experiment, participants had to perform several tasks involving prospective memory (PM) while their neural signal and their behavioural performance in tasks was recorded. In this study, both types of data were analysed to extract relevant characteristics (features) that were later used to discriminate between ASD and TD subjects on a single subject level. In Figure 2.1, the main stages, from experimental data collection to classification analysis, are presented. It is important to note that in this study 3 independent analyses were performed. First, a classification analysis was applied on the basis of their behavioural features. Subsequently, the same analysis was carried out using only features obtained from fNIRS data, and finally both types of features were integrated to perform a joined classification analysis. In this way, the discriminant power of both neuroimaging and behavioural information was first studied separately, and then together.

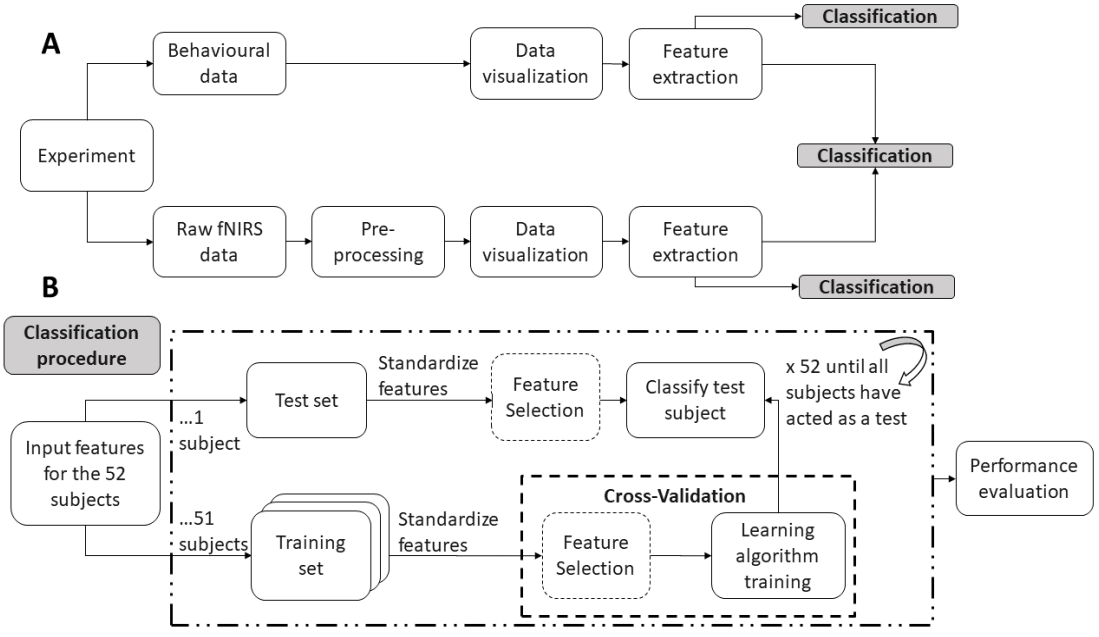



FIGURE 2.1: Flow diagram with the main steps performed on this study. Panel A shows the processes from experimental data collection until classification analysis. Panel B expands the stages involved on the classification stage represented as grey blocks in panel A.

Details of the steps shown in Figure 2.1 are described in this section. However, it is noteworthy that data collection and pre-processing steps were carried out in previous

studies, and therefore only the essential aspects of these steps will be described in this report. More details on the experimental design and fNIRS recording and pre-processing can be found in the report of the $3^{rd}$ mini-project [15].

## 2.1 Experimental Design

The present study analysed data from 26 TD (31.9±12 years) and 26 (33.2±10 years) participants with high-functioning ASD. The experimental protocol was divided in 4 ongoing (OG) and 8 PM blocks (Figure 2.3). During the OG blocks, participants were asked to respond whether the result of a mathematical operation was an even or odd number and to guess if the price of different items was above or below 14 (OG tasks). The answer was given by pressing the right and left arrow keys on a computer keyboard, and for each OG trial, their response time and accuracy were recorded.

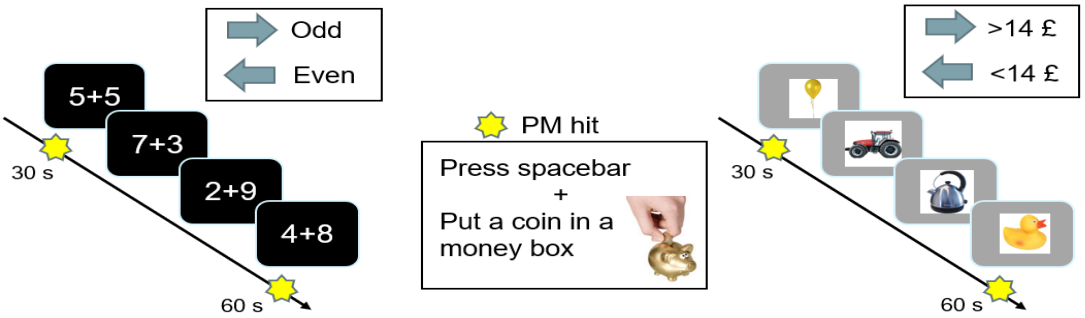



FIGURE 2.2: Illustration of the OG and PM tasks simultaneously performed during the PM blocks.

During the PM blocks, in addition to carrying out the OG task, they were asked to perform another task involving the prospective memory (PM task). For the PM task, subjects had to press the spacebar of the computer keyboard and had to drop a coin into a money box when they thought that 30 seconds had passed (Figure 2.2). The spacebar presses will be referred as PM hits.

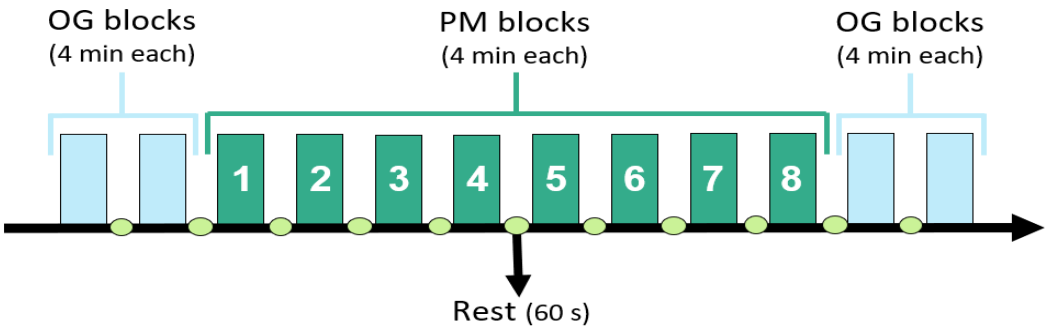



FIGURE 2.3: Schema of the different blocks in the experimental design.

In the different PM blocks, the tasks were carried out under different environmental scenarios. In some blocks, the coins could either be earned by the participants themselves or for others (Self/Other), and the experimenter could be present in the room or not (Present/Absent). The specific conditions of each block are illustrated in Figure 2.4.

| Experimenter | Earning money for: Other | Self |
|---|---|---|
| Present | 1, 7 | 4, 6 |
| Absent | 2, 8 | 3, 5 |

FIGURE 2.4: Design matrix describing the social conditions of each block.

## 2.2 fNIRS data acquisition and pre-processing

The hemodynamic response was measured over the PFC with the Hitachi WOT system [20], consisting on 6 light emitters (5 Hz sampling frequency) and 6 source detectors disposed in alternating fashion. As shown in Figure 2.5, 16 measurement points (channels) were placed at the mid distance between emitters and detectors. Each source emitted light at two different wavelengths (705 and 830 nm) to account for changes in both $HbO_2$ and HHb. Variations in light intensity were then converted to changes in optical density (OD) by means of the MBLL [13]. Then, OD data was pre-processed using Homer2 software package [21] to remove potential physiological noise and motion artifacts as described in Pinti et al., 2017 [22]. Finally, variations in OD were converted to concentration changes of $HbO_2$ and HHb according to Equation 1.1.

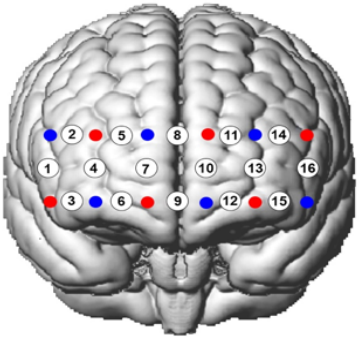



FIGURE 2.5: Configuration of the 16 channels in the fNIRS probe. The red and blue dots represent the position of light emitters and detectors respectively.

## 2.3 Feature extraction

In this study, data from the behavioural experiments and the fNIRS recordings was analysed to extract and visualize information that can best discriminate TD from ASD subjects. First, summary measures (features) of different aspects of both datasets were computed. A feature represents an interesting part of the original data that has been

processed and reduced into a single value. The goal of extracting features is to transform the information directly obtained in the experimental measurements into a more concise space in which to process the data. Once features were extracted, they were used as an input for a learning algorithm (Figure 2.1), which aims to recognise characteristic patterns of each group (TD and ASD) and detect important relations between features in a highly dimensional space. Then, the trained algorithm was used to predict the condition of a new individual (test subject) based on the same features. This approach enables to simultaneously look at the cross-information contained in different cognitive tasks and to integrate data from different modalities. This section describes the features extracted from the behavioural and fNIRS data as well as the procedure behind.

### 2.3.1 Behavioural features

During the experiment, participants were prompted to respond to the OG task as quickly and accurately as possible, and their response time (RT) to the stimulus and the accuracy of their answers were recorded. In addition, the time points at which subjects pressed the space bar (PM hits) were also analysed. In Table 2.1 the features extracted from the participants' behavioural data were summarized. For the extraction of some features, behavioural information was first analysed within each of the 8 blocks presented in Figure 2.3 with varying social conditions. Measures across blocks were subsequently computed. This approach was performed under the hypothesis that ASD participants might react differently (generally less mindfully) to the social manipulations of the experiment and different behavioural responses might be observed. First of all, their mean RT was computed as a feature (mRT). When visualizing the sequential RT of all participants throughout the experiment, different patterns were observed across blocks (Section 3.1.1). Therefore, the variance of RT in each block was calculated, and the mean (mBV) and variance (vBV) of this measure across blocks were included as features. As observed in Figure 2.6, in some cases, the evolution of RT during the first part of the experiment followed a different trend compared to the second half. This is particularly interesting as from block 1 to 4 participants were exposed to new environmental conditions (never faced before),

Table 2.1: Description of the behavioural features and their ID.

| Behavioural features | ID |
|---|---|
| Mean RT | mRT |
| Mean intra-block variance | mIBV |
| Variance intra-block variance | vIBV |
| Slope trend block 1-4 | $SLP_1$ |
| Slope trend block 5-8 | $SLP_2$ |
| Root-mean-square Deviation | RMSE |
| Mean PM interval | mPM |
| Variance PM interval | vPM |
| OG Accuracy | ACC |

whereas blocks 5-8 were conditions previously exposed in the experiment. For example, blocks 1 and 7 were carried out under the same scenario (Figure 2.4). In order to study whether this difference in trend varied across conditions, two linear models were fitted to the trials of RTs in the first and second part of the experiment. Then, the slope of the first ($SLP_1$) and second ($SLP_2$) linear trends were used as features. In addition, the root-mean-square deviation (RMSE) from the linear trend was computed for both lines and the mean value between the two was also included as a feature.

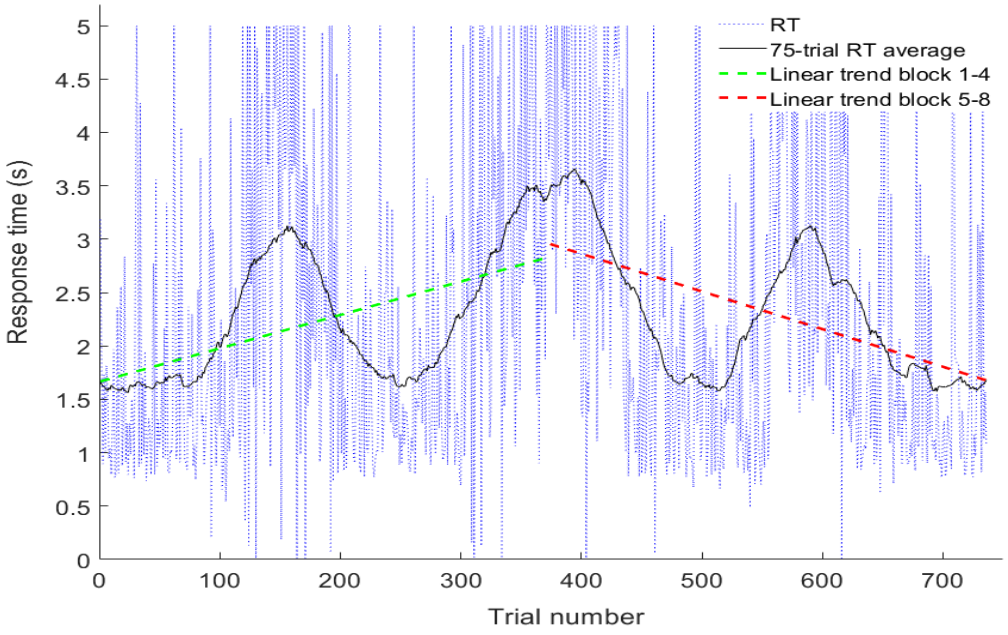



FIGURE 2.6: Example of the evolution of response times (RT) to the ongoing (OG) task throughout the experiment of one participant. Trials were plotted in sequential order, and a linear trend was fitted to the first and second half of the experiment. A 75-trial average was plotted to visualize the RT trend.

As previously mentioned, subjects were asked to press the spacebar every time they thought that 30 seconds passed. Therefore, the mean interval between all PM hits (mPM) and the variance of this interval (vPM) were both used to study how accurately and consistently the participants guessed the 30s time interval. All together, a total number of 9 behavioural features were considered for the classification analysis.

### 2.3.2 fNIRS features

fNIRS recordings were simultaneously measured by means of 16 different channels along the PFC (Figure 2.5), and concentration changes of $HbO_2$ and HHb were recorded throughout the experiment. Different features were then extracted from these measurements and then used as an input for the learning algorithm. However, it is important

to note that when dealing with small datasets, using a large number of input features for the algorithm can worsen its performance, as more predictors would require more parameters to be estimated. Therefore, channels were grouped into 4 main regions. The fNIRS data considered in this study was exclusively extracted from the channels within these regions. Features were first extracted for each of the channels considered, and then they were averaged within regions. As observed in Figure 2.7, the regions and channels considered included: R1 (channels 8, 9), R2 (channels 7, 10), R3 (channels 14, 15, 16) and R4 (channels 1, 2, 3).

These regions were selected in order to preserve spatial information and reduce the dimensionality of the data. According to the PM tasks and social conditions, similar activation patterns were expected within the averaged channels. First, central regions (R1 and R2) were selected to account for neural activation during PM tasks, as increases in regional cerebral blood flow during PM paradigms have been mostly detected in the rostral PFC [23]. In particular, pairs of channels 8-9 and 7-10 were grouped together since homologous regions in each hemisphere of the rostral PFC have generally shown to co-activate [24, 25]. In Catani 2012 [26], a strong connectivity between populations of neurons in R3 and R4 was observed. Therefore, channels 1-2-3 and 14-15-16 were also averaged to account for lateral activation.

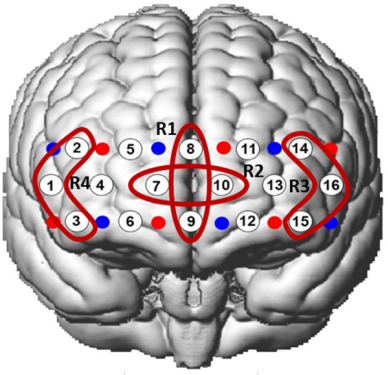



FIGURE 2.7: Illustration of the channels present in each of the 4 regions considered in this study.

All features extracted from the fNIRS data are shown in Table 2.2. Due to their physiological relation, anti-correlation between $HbO_2$ and HHb is often observed and expected during fNIRS recordings. However, if certain conditions (e.g. ASD) present different patterns of neural activation, the degree to which both signals are anti-correlated can differ. Hence, the first feature extracted was the Pearson correlation coefficient between $HbO_2$ and HHb ($R_{1-4}^{HbO_2-HHb}$). The second type of feature considered were the $\beta$-values obtained when applying the general linear model (GLM, [27]). Before applying the GLM, the correlation-based signal improvement method was applied (CBSI, [28]). This transformation was obtained by a weighted linear combination between $HbO_2$ and HHb, which maximizes their anti-correlation, removing any residual motion artifacts and providing a more robust detection of functional events [22]. Then, the GLM was applied to the CBSI signal:

$$Y = X\beta + \epsilon \tag{2.1}$$

Where Y is the matrix containing the CBSI signals from all channels (each column corresponds to a channel), X is the design matrix with the predictors, $\epsilon$ is an error term matrix representing the residual variance, and finally $\beta$ is a matrix containing the coefficients, which indicate the contribution of each predictor to the response variable Y. At this stage, the PM blocks were modelled as 8 - 4 minute blocks in the design matrix. The contrast between earning for themselves vs earning for others (blocks 3, 4, 5, 6 vs 1, 2, 7 and 8) was computed to test the effect of the reward. Then, $\beta$-values resulting from this contrast were used as input features for the algorithm. This comparison Self vs Others was performed in previous studies and it was selected in this case since significant differences between groups and significant brain activity within groups had been exhibited.

In previous studies with this dataset, the signal contained from 5 seconds before to 5 seconds after the PM hits was extracted and analysed, as it is in this time window where most neural activation patterns in response to the PM task were expected [15]. However, recovering the real onset of functional events (neural activity in response to a functional task) during PM tasks can be challenging, since the PM intention is maintained over long periods in participants' mind and the moment of the recall of the intention can be hard to predict. Therefore, in this report we aimed to detect activation patterns through a recently developed method: the Automatic IDentification of functional Events (AIDE, [22]) algorithm for fNIRS signals. This method aims to detect functional events by convolving a boxcar function with the canonical hemodynamic response function (HRF) to generate an activation model that will be later fit to the CBSI signal by means of the GLM. Figure 2.8 illustrates the performance of the algorithm at detecting functional events from the CBSI signal. For details of this method see Pinti et al., 2017 [22]. As an output, the algorithm returns the time location and duration of functional

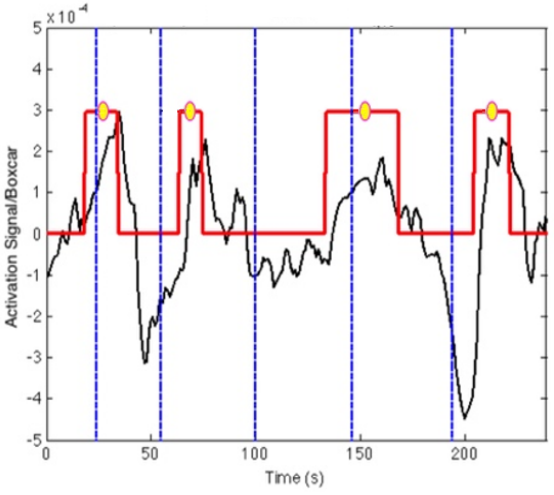



FIGURE 2.8: Example of the AIDE model (–) applied to the CBSI signal (–) of one participant. The yellow dots are located at the mid-point of the functional event. Temporal locations of the PM hits are represented with a vertical blue dashed line (- -).

events. The algorithm was applied to the fNIRS data and a number of features regarding functional activation patterns of the participants were extracted. First, the number of events throughout the whole experiment (NE) was computed for each region (1-4) in order to detect whether a different frequency of functional activation was present between the groups. Then, the median and interquartile range (IQR) of the events' duration were used as features (mDE and vDE). It is important to note that median and IQR were used instead of mean and variance. The main reason behind this decision was to avoid the effect of outliers caused by miss-detection of events. Subsequently, the relation between the PM task and the corresponding functional activation was analysed by computing the distance between the mid-point of the detected functional event and the nearest PM hit (see Figure 2.8). For this computation the time (in seconds) of the mid-point of each event was extracted, and the timing of the nearest PM hit was subtracted from this value. In this way, this distance could take either positive or negative values depending whether the nearest PM hit occurred before or after the functional event. The median and IQR distance event - PM hit (mE-PM and vE-PM) were both extracted only for R1 and R2, as these were the areas where neural activation related to the PM task was expected. The distance between consecutive functional events was also obtained by extracting all the time intervals between the mid-points of consecutive events and performing the median and IQR (mIE and vIE).

TABLE 2.2: Description of the fNIRS features, ID and regions considered in each case.

| $N^0$ | fNIRS features | ID | Regions included |
|---|---|---|---|
| 1-4 | Correlation $HbO_2$-HHb | $R_{1-4}^{HbO_2-HHb}$ | 1 to 4 |
| 5-8 | Beta values | $\beta_{1-4}$ | 1 to 4 |
| 9-12 | Number events | $NE_{1-4}$ | 1 to 4 |
| 13-16 | Median duration events | $mDE_{1-4}$ | 1 to 4 |
| 17-20 | IQR duration events | $vDE_{1-4}$ | 1 to 4 |
| 21-22 | Median Distance Event - PM hit | $mE\text{-}PM_{1-2}$ | 1 to 2 |
| 23-24 | IQR Distance Event - PM hit | $vE\text{-}PM_{1-2}$ | 1 to 2 |
| 25-28 | Median interval between events | $mIE_{1-4}$ | 1 to 4 |
| 20-32 | IQR interval between events | $vIE_{1-4}$ | 1 to 4 |
| 33-38 | Correlation between regions | $R_{regionA-regionB}$ | 1 to 4 |

Finally, the last fNIRS features extracted analysed the correlation between regions (i.e. functional connectivity). A simple Pearson correlation coefficient was obtained by first averaging the signal of the total haemoglobin ($HbO_2$ + HHb) within same-region channels. Then, the correlation between the mean signal in each region was computed. A total number of 6 possible correlation coefficients between pairs of regions were computed ($R_{regionA-regionB}$), which constituted features 33 to 38 (Table 2.2). All together, a total number of 38 fNIRS features were included in the classification analysis.

## 2.4 Classification and feature selection

Once relevant features from the original data were extracted, these were transformed to standardized z-scores and used as an input for the classification analysis. Due to their general capacity to find patterns and successfully deal with multidimensional feature sets, a multivariate machine learning approach was used in this report to predict the condition of a subject (TD or ASD) given a number of input features. Machine learning classifiers have already been applied to develop diagnostic tools for a variety of brain disorders (e.g. Schizophrenia, depression, ADHD), and they have proved to be particularly suited to identify conditions at the individual level [16]. Since the condition of all participants was previously known in this study, the classification procedure belongs to the field of supervised learning.

Three separate analyses were performed in this report (Figure 2.1 panel A). The first one consisted in using the 9 extracted behavioural features to asses the discriminant capability of the learning algorithm. Analogously, the same classification analysis was performed using the 38 fNIRS features. Finally, behavioural and fNIRS features were combined and used as an input for the classifier. In Figure 2.1 panel B, the main steps involved in the classification analyses are presented. Most machine learning platforms dealing with a large number of observations involve the partition of the dataset into a training and test group. The training set is used to select the most relevant features and to train the classification algorithm. Then, the model developed is finally applied to the test set and the prediction capacity of the algorithm is assessed. However, when dealing with small number of observations the partition into training and test sets usually results in a very small training set size which is not sufficient to train the classifier. As a result, most neuroimaging studies aiming to predict conditions with a limited number of participants (which is often the case) apply other approaches such as random sub-sampling or leave-one-out (LOO) classification [19, 29]. For this study, the LOO approach was used, which consisted in splitting the dataset into 1 participant who formed the test set, and 51 subjects that were used to train the classifier (training set). The trained classifier was then used to decode the label (TD or ASD) of the tested subject. This process of splitting the dataset was repeated 52 times in turn, with a different participant acting as a test each time until all of them were classified.

For the analyses using fNIRS or behavioural features exclusively, classification with and without the feature selection step was compared. Feature selection consisted on a previous stage that selected the most relevant features to use them as an input for the classifier. In this section, details on the classification algorithm and feature selection are described.

### 2.4.1 LogitBoost algorithm

No machine learning algorithm has proven to be universally better than another, instead, different types of models have shown more suitability for certain contexts. One group of algorithms that have been particularly successful in a variety of applications, are boosting methods [30]. Boosting algorithms are ensembling methods, which combine a number of "weak" classifiers (or weak learners) and integrate them in a voting scheme to generate a powerful classification model that performs better than any of the individual weak learners alone [31]. In the context of supervised learning, a binary classification problem of sample size $n$ (number of subjects for this study) can be formulated as follows [17]:

$$X = \Big\{(x_i, y_i),\ x_i \,\epsilon\, R^d,\ y_i \,\epsilon\, \{-1, +1\},\ i = 1, 2, ..., n\Big\} \tag{2.2}$$

Where $x$ is the input feature matrix of size $n$ x $d$, $d$ is the number of features, and y is the response vector with the labelled classes (either -1 or 1). From a training set, the aim is to develop a classifier that decodes the binary state $y$ of a number of unseen test samples given their input features ($x$). In order to approach this problem, most machine learning algorithms develop a single classifier $h(x)$. However, ensembling methods develop not only one, but a group of classifiers $\{\ h_1(x),\ h_2(x)\ ,..,\ h_T(x)\ \}$, each of which will determine the labels of the input feature matrix. Then, an additive ensembler algorithm assigns a set of weights $\{\ \alpha_1,\ \alpha_2,\ ..,\ \alpha_T\}$ to the weak learners, and it develops the final classifier as a weighted average [17]:

$$H(x) = \ \alpha_1 h_1(x) + \ \alpha_2 h_2(x) + \ .. + \ \alpha_T h_T(x) \tag{2.3}$$

Depending on the score $H(x)$, the ensembler will determine the label of an observation as either positive (+1) if $H(x) > 0$ or negative class (-1) if $H(x) < 0$. The higher the absolute value of $H(x)$, the more likely the classified observation belongs to the predicted class.

The goal of boosting algorithms is to train the parameters from the weak learners $\mathrm{h}_{1..T}(x)$ and optimize the weights of the ensembler. Given a training set and an initial distribution of weights, boosting methods update the weights and parameters from the weak learners at each iteration of the algorithm. This process is then repeated until a certain number of iterations is reached. There are different ways in which the optimization of parameters can be performed, and the particular methodologies give rise to multiple versions of boosting methods [31].

For this study, the Logitboost algorithm was used. Generally, to fit the model parameters', most algorithms try to minimize a loss function. Logitboost was developed by Friedman et al., 2000 [31], and it minimizes a binomial log-likelihood loss function, which is linearly dependent to the misclassification rate and has proven to be more robust against noise and outliers than other boosting methods (e.g. AdaBoost) [32]. Given a boosting algorithm and a training set of $N$ observations, LogitBoost is developed by minimizing the logistic loss function:

$$L(y,\ H(x)) = \sum_{n=1}^{N} w_n\ log(1\ +\ e^{\ y_n H(x_n)}) \tag{2.4}$$

Where $y_n$ are the true labels, $w_n$ are the observation weights and $H(x_n)$ are the predicted scores from the classifier [32].

One of the key features that characterize boosting classifiers are the weak learners. Different types of classification models can be used to constitute the weak learners of an ensembler (e.g. $k$-nearest neighbours, support vector machines, discriminant analysis etc.). However, due to their flexibility and capacity to deal with high dimension datasets, decision trees are the most common weak learners in boosting methods [33]. For binary classification, given a feature matrix $x$ with $d$ predictors, decision trees first try to split the data into 2 groups and generate the first node. The splitting node represents the value of a certain predictor (feature) that maximizes the separation between the two classes according to a certain criterion (e.g. gini index [34]).

Then, a new node is formed by testing another splitting point in the feature space, and this process is repeated a certain number of times. A new sample is then classified by first testing the value of a certain feature at the first node, and then moving down the tree branch until reaching the bottom of the tree, which determines the specific label of that sample [35]. In Figure 2.9 an illustration of a decision tree with 3 nodes to decide whether a day is suitable for playing tennis can be observed. The outputs of the tree (Yes/No) are located at the bottom, after the value of a number of attributes (Outlook, humidity and wind) were tested. For more details about the parameters of decision trees see Roe et.al., 2006 [34].

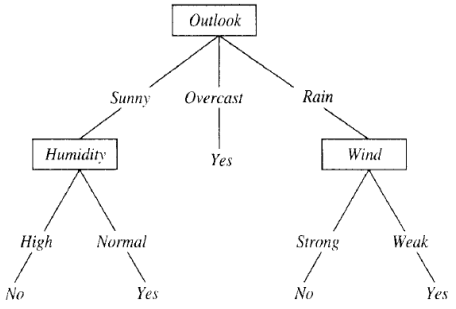



FIGURE 2.9: Example of a decision tree to predict whether a certain day is suitable for playing tennis depending on the value of certain attributes. Figure from [35].

As previously explained, the learning algorithm aims to optimize a number of parameters during the training stage to build the classification model. However, some algorithms contain a group of specific parameters that need to be specified before the learning process starts, and these are referred to as hyperparameters. The methodology and rationale employed to determine the values of the hyperparameters depends on the specific problem, the expertise of the model developer and the computational time required to optimize those. Some common methodologies used in the machine learning community include genetic algorithms, random-search, Bayesian optimization or grid search [36]. On the other hand, it is also common to find studies that merely used the default hyperparameters settled in a specific toolbox [17, 19, 29]. Nevertheless, optimizing the hyperparameters of a machine learning algorithm can significantly improve the models' performance [16].

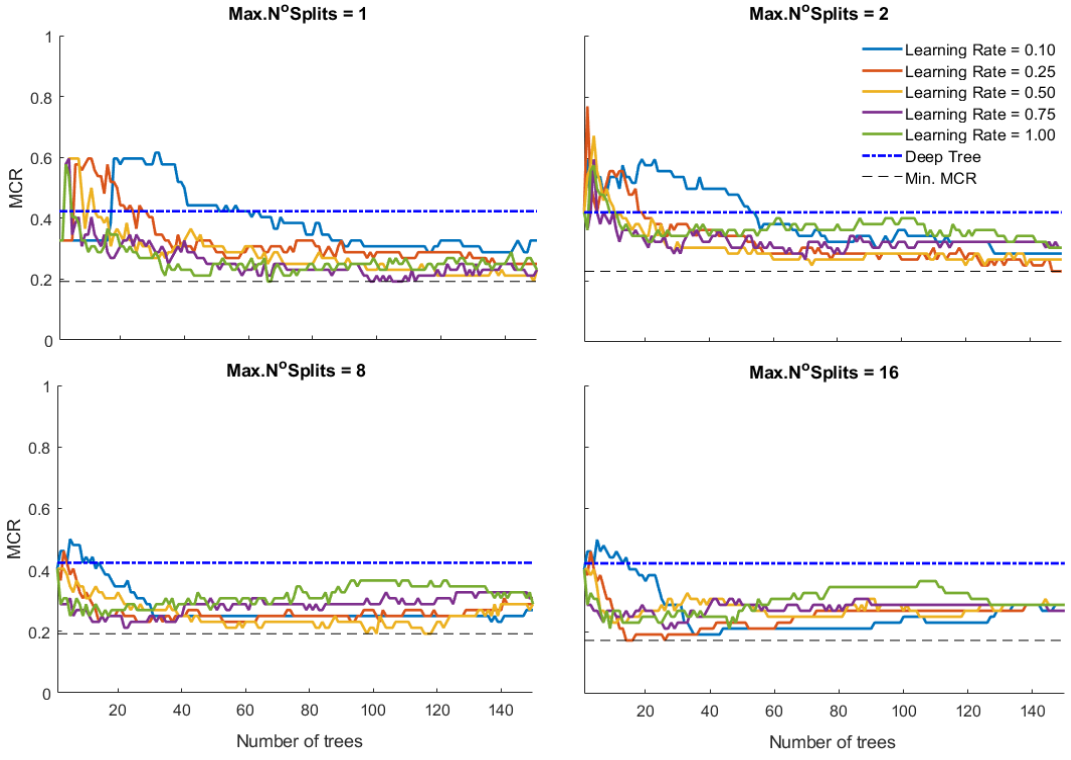



FIGURE 2.10: Example of grid search for hyperparameter optimization. Misclassification rate (MCR) in the test set was measured during leave-one-out cross-validation when varying learning rate, maximum number of splits and number of trees in the LogitBoost algorithm. The combination of hyperparameters that minimized MCR was selected to train the model.

The main 3 hyperparameters to optimize in LogitBoost are: (1) the number of weak learners (for our study number of decision trees), (2) the maximum number of splits that each decision tree can generate and (3) the learning rate. The learning rate can take values from 0 to 1, and in the case of boosting it limits the number of boosting iterations that the algorithm performs [31]. In this study, the method implemented to

tune the hyperparameters of the model given a training set was grid search. For the implementation of the grid search methodology, a number of candidate values for each hyperparameter was initially generated, forming a grid with combinations of different values for the 3 hyperparameters to be optimized. Then, for each possible combination of hyperparameters the classification performance was assessed using a leave-one-out cross-validation approach (LOOCV).

For each combination of hyperparameters, the model was trained using all the observations except one (test), and the left-out sample was classified. This process was then repeated in turn, until all subjects were classified when acting as a test. Finally, the misclassification rate (MCR) was used to assess the performance of the particular combination of hyperparameter values. For this study, values of 0.1, 0.25 0.5 0.75 and 1 were studied for the learning rate. The number of trees was increased from 1 to 150, and the maximum number of splits was varied from $2^0$ to $2^5$. Figure 2.10 illustrates the hyperparameter optimization performed during one of the classification analysis. MCR was plotted as a function of the hyperparameter values. In this case, all results were compared to the ones of a single deep (no maximum number of splits specified) decision tree. From the example shown in Figure 2.10, the combination of hyperparameters that minimized misclassification rate was found at 14 week learners with a maximum of 16 number of splits and a learning rate of 0.25.

For the implementation of the boosting algorithm, the platform *fitcensemble* from MATLAB was used [37].

### 2.4.2 Feature selection

One common and highly important step on most classification platforms is the implementation of a feature selection stage. Even with machine learning algorithms that are capable of dealing with high dimensional spaces, irrelevant features can reduce the performance of the classifiers [29], specially when the sample size is limited. In classification tasks, feature selection aims to find a subset of predictors that describe the most informative characteristic patterns from the classes and omit irrelevant information. For this study, the implementation of feature selection was performed for the behavioural and fNIRS analyses and compared to the procedure without this stage. It is important to note that feature selection should only be performed in the training set in order to avoid overfitting [16], given that when using the complete dataset to select the most relevant features the algorithm might find a discriminant pattern in a specific dimensional space, but this needs to be contrasted with a validation set to ensure that the discriminance found is not the result of sample noise (overfitting). Therefore, for each of the 52 folds

performed during the classification analyses the most relevant features were selected from the training set.

There are two main types of feature selection approaches: filtering and wrapper methods. Filtering methods have been widely used on the neuroimaging community and they look at the general statistics of the features without involving any learning algorithm. A common filter approach is to apply univariate group-level statistical tests. Generally, analysis such as t-tests are used to compute the p-values of each feature at the group level. Then, p-values are used as an index to score the discriminant power of each feature and to select those features with the most strongly significant p-values as an input for the classification analysis. Despite being computationally fast, these univariate approaches assume that there is no interaction between features. Assuming a direct relation between the p-values and the discrimination power does not always result in the best results, as some features that are individually irrelevant might provide crucial information when used in combination with others. Therefore, selecting features based on univariate tests sensitive to group mean can result in the loss of valuable information [16].

On the other hand, wrapper methods implement the learning algorithm to assess the predictive power of different features when used in combination [16]. However, given a number of $d$ features, $2^d$-1 possible combinations of feature subsets can be generated, and an exhaustive analysis over all possible combinations is usually not feasible, as for each feature subset the learning algorithm needs to be trained and evaluated. As a result, different search methods have been developed to solve this problem. Here, a sequential forward selection (SFS) method was applied, which is one of the most common approaches in wrapping algorithms. This method was used with a LOOCV approach to obtain the MCR that results from applying the learning algorithm with a specified combination of features. MCR was used as a criterion to find the best feature subset. SFS starts by evaluating MCR for each feature independently and the feature that minimizes MCR is added to an empty candidate set. Then, the algorithm evaluates the effect of adding each of the remaining features to the candidate set, and the one that minimizes MCR is added. This process is repeated until the addition of a new feature does not decrease MCR and the optimal set is then generated. This has proven to be a particularly efficient search method, but one of the main drawbacks is that the algorithm might stop prematurely in the presence of a local minimum in the MCR function. For example, in Figure 2.11b, it can be observed that the number of features that minimizes MCR is found at 7 features. However, the fact that MCR does not improve from 4 to 5 features would make the SFS algorithm stop at 4 and not selecting the most efficient subset.

To tackle this issue, it is also possible to specify the number of features to select, and the algorithm sequentially adds the most relevant features until reaching the specified number. In this study, SFS was first applied over the whole set, specifying a number of features to select equal to the feature set ($d$). In this way MCR was plotted as function of the number of features (Figure 2.11), and the specific number of features that minimized MCR was observed for the behavioural and fNIRS analyses. Thus, the optimal number of relevant features was elucidated from the MCR plot.

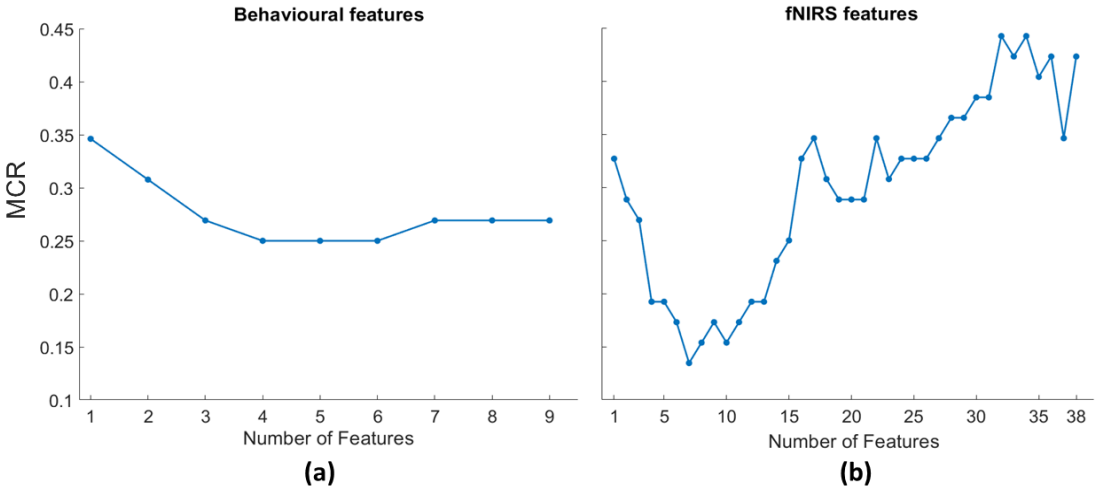



FIGURE 2.11: Analysis to find the optimal number of features that the wrapper method selected. Misclassification rate (MCR) during leave-one-out cross-validation was plotted as a function of the number of features that the wrapper method included in the classification. The optimal number of features for the behavioural (a) and fNIRS (b) sets was found at 4 and 7 respectively.

Once the optimal number of features over the whole set was identified, at each one of the 52 iterations involved in the classification analysis (Figure 2.1), the wrapped method was required to sequentially add features until reaching the specified optimal number. The results from the optimal number of features analysis can be observed in Figure 2.11. It is important to note that MCR exhibits really low values in this plot. This can be explained by the fact that in this procedure all observations were included, and therefore the parameters of the model were optimized over the whole set, generating really low MCRs. However, this was only performed to establish the optimal number of features that the SFS was required to add in each fold, but the performance during the LOO classification analysis was later assessed. As a result from these plots, 4 and 7 features were selected at each fold of the behavioural and fNIRS classification analyses respectively. Importantly, one crucial fact that can be observed from this graph is that in both cases, the addition of more than one variable improved MCR, which points out the benefits of multivariate approaches for classification at the subject level against univarite techniques.

## 2.5 Performance evaluation

In this report, the ASD and TD groups were labelled as the positive and negative class respectively. As previously mentioned, one subject was classified at a time while all the others were used to train the model, and at the end of this leave-one-out procedure, all subjects (N) were labelled by the algorithm. Once all participants had been decoded, to evaluate the performance of the model at discriminating between TD and ASD, 4 measures were first computed and the subjects were divided into:

1. True Positives (TP): number of ASD subjects correctly labelled.

2. True Negatives (TN): number of TD subjects correctly labelled.

3. False Positives (FP): number of TD subjects that were misclassified as ASD.

4. False Negatives (FN): number ASD subjects that were misclassified as TD.

Then, accuracy, specificity and sensitivity were used to quantify models' performance:

$$Accuracy = \frac{TP + TN}{N} \tag{2.5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.7}$$

In addition, another common metric for binary classification tasks was used, which is the area under the curve (AUC) from the receiver operating characteristic curve (ROC curve). The ROC curve illustrates the true positive rate (TPR, sensitivity) against the false positive rate (FPR, 1 - specificity) when the discrimination threshold of a classifier is varied (Figure 2.12). ROC curves are widely used in a variety of fields to compare the performance of several classifiers by analysing the AUC. Some of the main advantages of the AUC is that it does not assume any probability density function of the model and it is independent from the prior probabilities of each class [38]. In addition, the AUC is not biased by the selection of certain thresholds that might be decided by experts and is therefore considered an intrinsic and robust metric for the performance of a classifier. The performance of classification models is normally compared to the one of a random classifier in the ROC graph (dashed black line Figure 2.12), which would have an AUC of 0.5.

As illustrated, an ideal classifier would have a constant TPR of 1, independently of the FPR and an AUC of 1. One of the main advantages of computing the ROC curve is that it enables decision makers to identify the best operating point, which is the optimal point that minimizes the balance between the misclassification of true classes and the number of false positives. In Figure 2.12, the optimal operating point of a certain classifier (A) is shown by a red circle. In this report, for each classification analysis, ROC curves were computed from the scores of the classifier. As previously mentioned, for each instance, LogitBoost returns not only the predicted class but also a score that indicates the confidence that a particular observation belongs to the predicted class. The score returned by the algorithm can be understood as the likelihood that the classified instance comes from the positive class (posterior probability), and the relation between scores and accurate predictions is used to asses the classifier and to plot the ROC curve. Overall, for each classification analysis performed in this study, accuracy, sensitivity, specificity and AUC were reported.

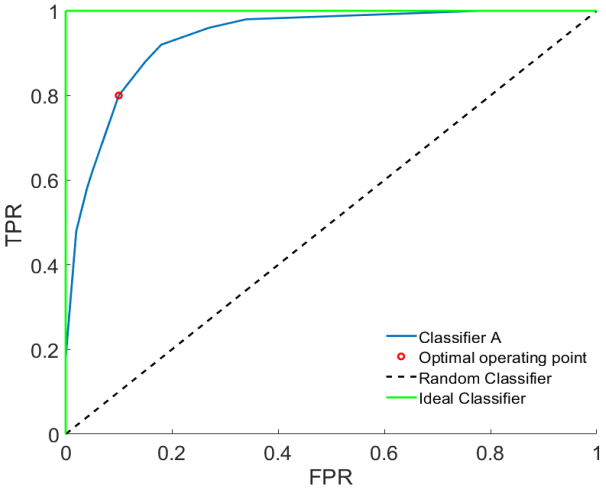



FIGURE 2.12: Example of a receiver operating characteristic (ROC) curve. The performances of a random (- -), unknown example (—) and ideal (—) classifiers are illustrated.

# Results

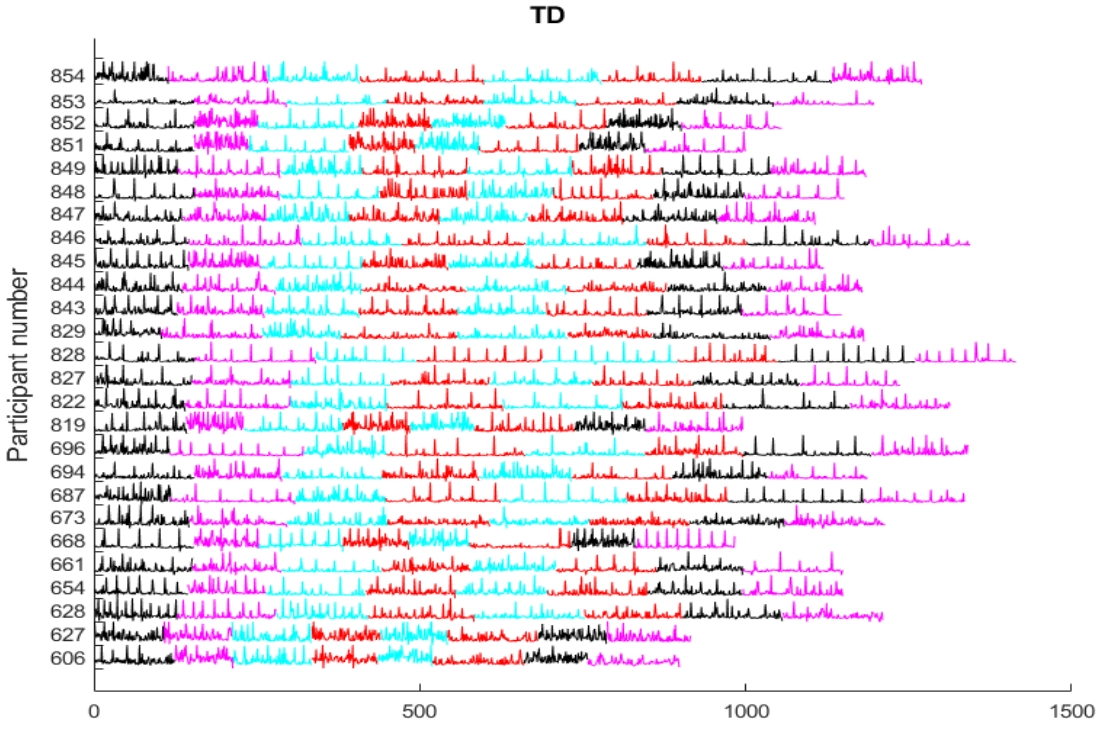
The results from this study were divided in two main sections. First, behavioural and neuroimaging data were separately analysed, and visualization techniques and summary statistics were applied to detect trends and characteristics patterns in the data. On the second part, results from the classification analysis were presented, first for behavioural and neuroimaging features separately and finally for the integrated analysis.

## 3.1 Data visualization and summary statistics

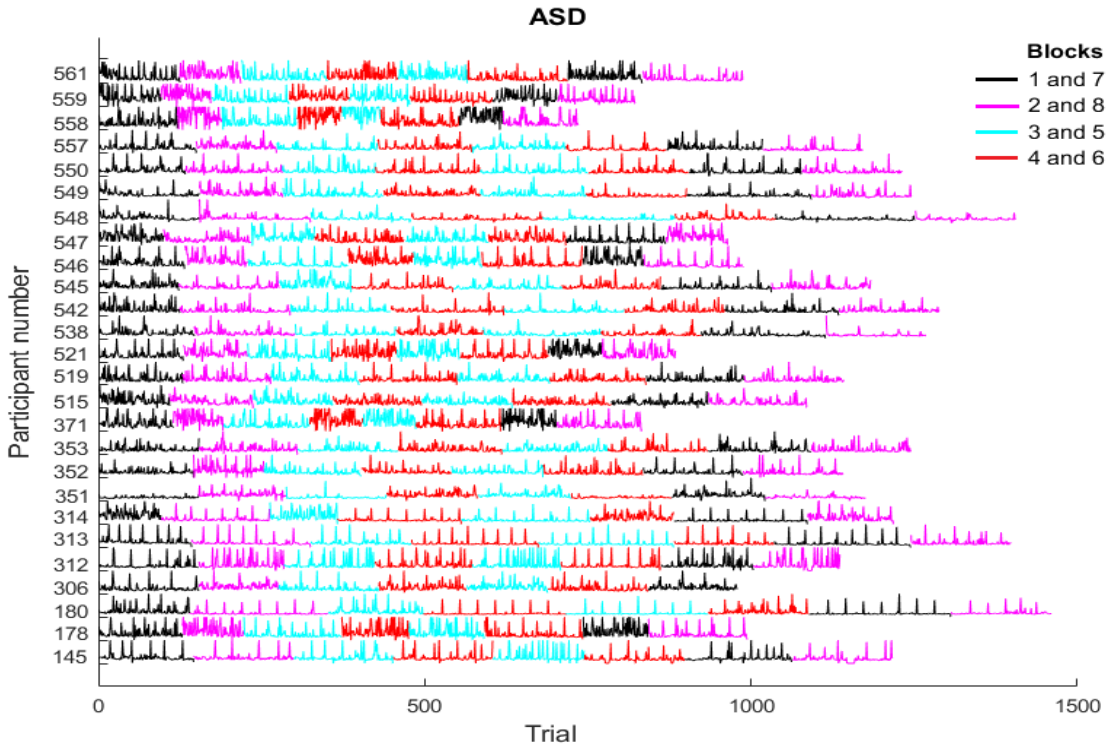
Data visualization represents a key stage on most classification platforms. Visualizing raw data and engineered features can have great advantages. First, it can help on detecting meaningful patterns and relations between variables before any classification algorithm is applied. Moreover, it can be used to identify outliers, visualize the spread of the data and detect central tendencies. But most importantly, when aiming to detect functional biomarkers for a certain disease it can have a key role on understanding the outputs from classification results, indicating why some features might be more relevant than others and helping on the intepretability of relations build inside the model.

### 3.1.1 Behavioural features

First, response times (RT) in each trial across the experiment where plotted for each participant. Figure 3.1 shows the evolution of RTs for the TD and ASD groups separately, and coloured according to each block. Each subject was plotted separately (one above the other) in order to visualize patterns in the whole group and compare across conditions. The RTs of each participant were separated with a 5 seconds range. In this way, each tick in the y axes represents the 0 s reference point for one subject and the 5s limit for the previous one. In Figure 2.6 the evolution for a single subject (561) was presented. Subsequently, the slopes $SLP_1$ and $SLP_2$ (Section 2.3.1) were plotted one against the other to study differences in the linear trend between the first and second part of the experiment (Figure 3.2). It is noteworthy that each of the 2-D square spaces generated in that figure represents either a change (from increasing to reducing RT or vice-versa) or a consistent trend throughout the experiment (either reducing or increasing RT).




(a)




(b)

FIGURE 3.1: Evolution of response time (RT) to the ongoing task (OG) for TD (a) and ASD (b) participants. The RTs of each subject were plot with a separation of 5 seconds and each tick in the y axes represents the reference point (0s) for each participant. RTs were plot for each trial in sequential order and coloured according to each block.

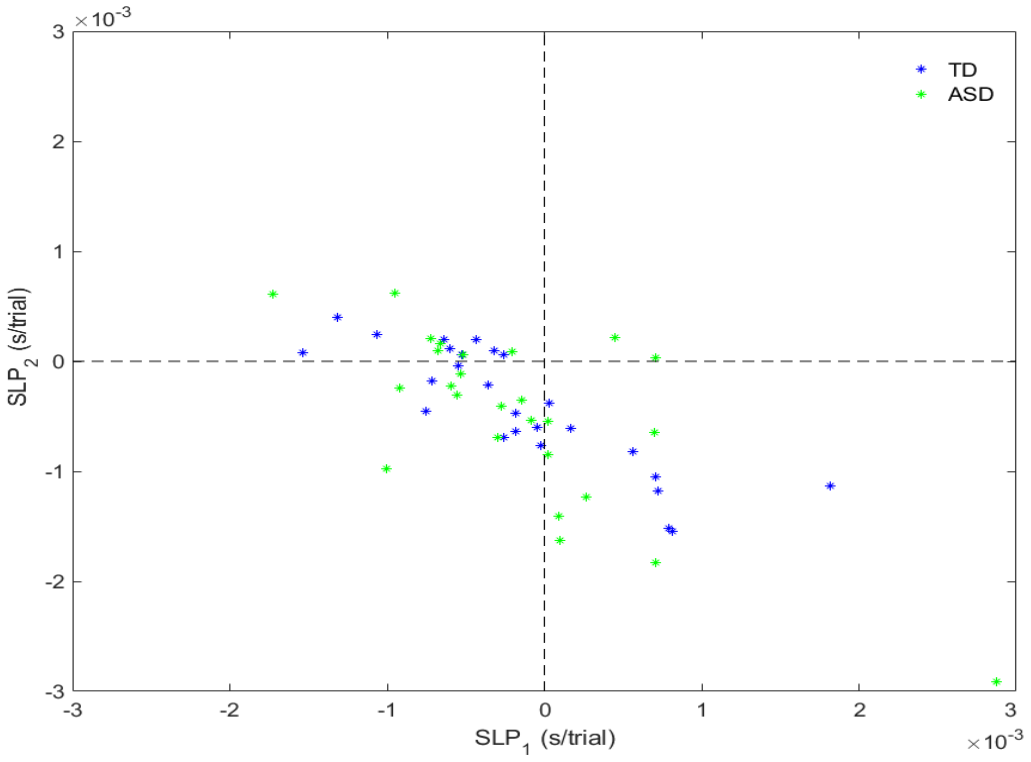



FIGURE 3.2: Scatter plot comparing the slope from the linear trend of the response times (RTs) in the first half of the experiment ($SLP_1$) against the second half ($SLP_2$).

Figure 3.3 shows the scatter plot for the features mIBV vs vIBV. This was illustrated to exhibit the distinct linear trend that was observed on this space between both groups when a linear model was fitted. The labels of each individual were also included in the figure in order to compare the results from the scatter plot with Figure 3.1.

In Figure 3.4, the interval (in seconds) between each of the PM hits was plotted per participant. Due to the fact that each subject was asked to press the space-bar every 30 seconds in all of the blocks, all PM hits were considered for this plot and the optimal 30 seconds interval line was plotted to visualize the dispersion from this optimal performance.

The accuracies of the OG tasks were then computed. During each of the OG trials, subjects could either answer correctly or incorrectly (1/0), and the average performance was computed. Then, a boxplot from the mean accuracies of each subject was generated in Figure 3.5, separating by condition. The limits of each box constrained the $25^{th}$ and $75^{th}$ percentiles, while the red-mark exhibited the median of each condition. Whiskers were extended until the extreme points not considered outliers.

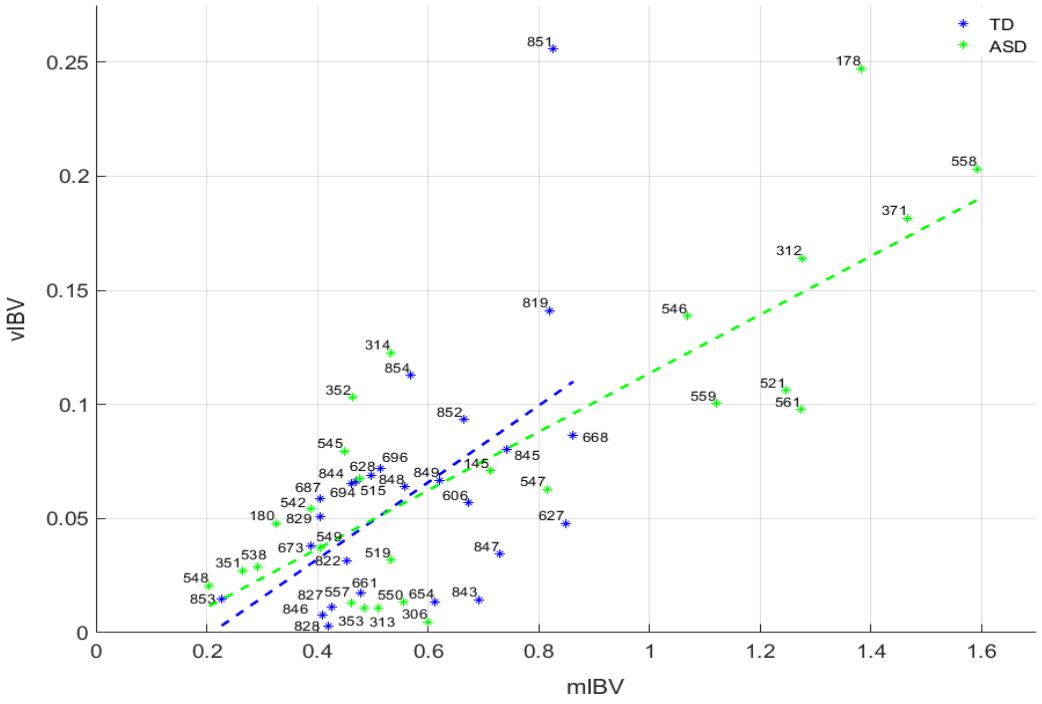



FIGURE 3.3: Figure comparing the mean variance between blocks (mIBV) against the variance of intra-block variances (vIBV). Linear trends were fitted for both TD (blue) and ASD (green). Labels represent the identifier of each participant during the experiment.

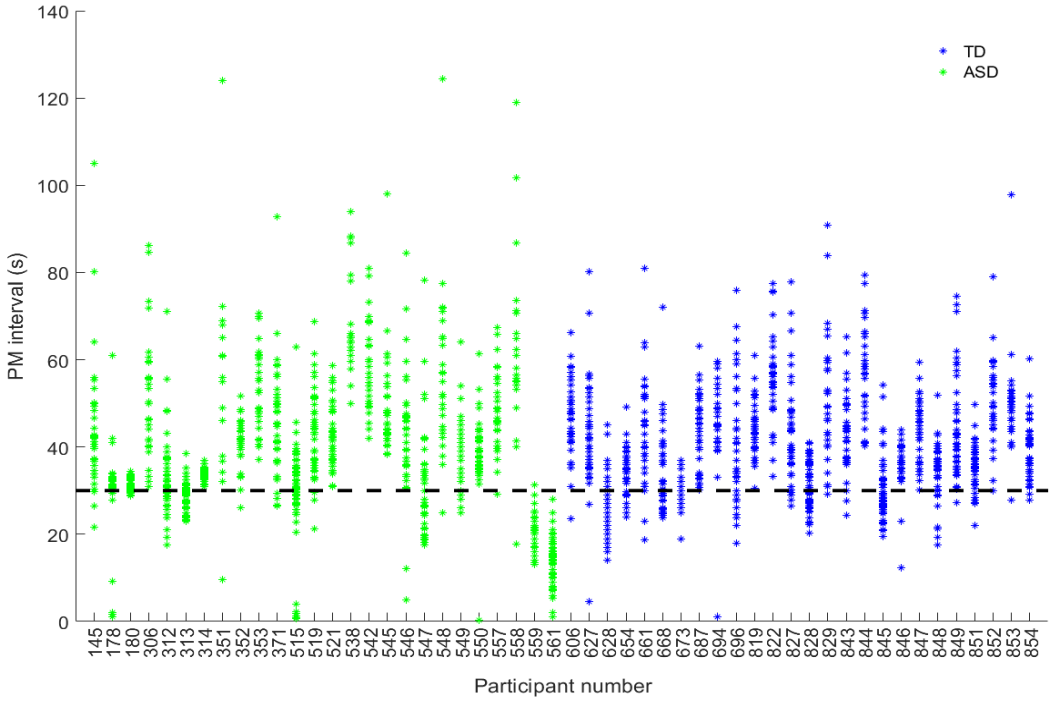



FIGURE 3.4: Prospective memory (PM) task time intervals were plotted for each subject. Participants were asked to press the space-bar every 30 seconds. Hence, this optimal interval was plotted with a dashed line.

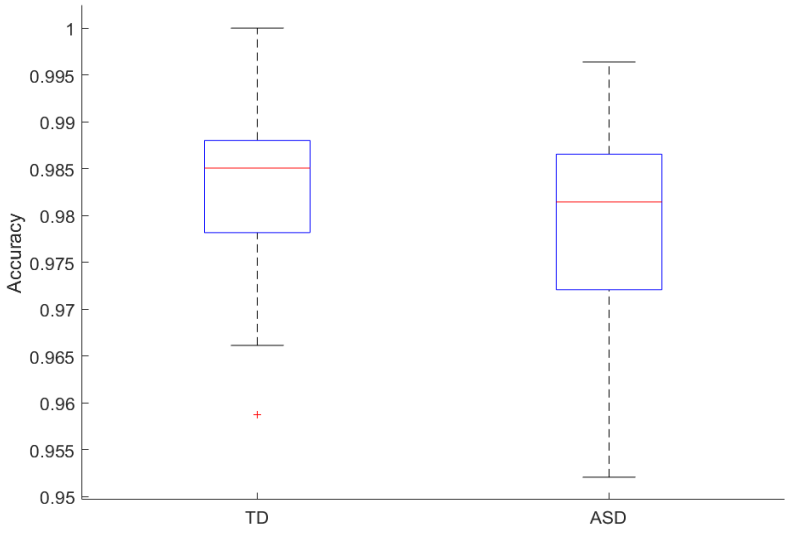



FIGURE 3.5: Boxplot showing the distribution of accuracies during the ongoing (OG) task depending on group condition. The mean accuracy between all OG trials was computed per subject. Due to the relative simplicity of the OG tasks all accuracies ranged above 95%.

Finally, a summary statistics table was generated. In order to analyse differences in central tendency between groups, the mean value of each feature for each condition was shown in Table 3.1 together with the standard deviation. Then, univariate t-test were applied to detect significant differences at the group level by means of two-sample t-tests. Since the t-tests were performed repeatedly across the 9 features the Bonferroni correction was applied, considering the significance level at $p$-value=0.05/9=0.006.

TABLE 3.1: Summary table for the 9 behavioural features. The mean of each feature is shown together with the standard deviation for each condition. Mean values for each feature were compared across conditions using a two-sample t-test. The $p$-value resulting from this test was computed. $p$-value <0.006 was considered a significant result. Features in bold were the ones later selected by the classification algorithm.

| N$^o$ | Feature ID | TD | ASD | $p$-value |
|---|---|---|---|---|
| 1 | mRT | 1.29 ± 0.21 | 1.38 ± 0.34 | 0.266 |
| 2 | **mIBV** | 0.56 ± 0.16 | 0.72 ± 0.42 | **0.275** |
| 3 | vIBV | 0.06 ± 0.05 | 0.07 ± 0.06 | 0.970 |
| 4 | **SLP$_1$** | -1.59 $10^{-4}$± 7.36 $10^{-4}$ | -1.51 $10^{-4}$± 8.49 $10^{-4}$ | **0.697** |
| 5 | SLP$_2$ | -4.17 $10^{-4}$± 5.59 $10^{-4}$ | -4.93 $10^{-4}$± 8.05 $10^{-4}$ | 0.677 |
| 6 | RMSE | 0.13 ± 0.07 | 0.13 ± 0.07 | 0.083 |
| 7 | mPM | 41.48 ± 8.07 | 41.70 ± 13.24 | 0.943 |
| 8 | **vPM** | 85.57 ± 55.7 | 135.04± 134.09 | **0.092** |
| 9 | **ACC** | 0.98 ± 0.01 | 0.97 ± 0.01 | **0.159** |

### 3.1.2 fNIRS features

Due to the large number of fNIRS features (38), a summary statistics table was computed to analyse differences in central tendency and dispersion by means of group mean and standard deviation. In Table 3.2, the mean and standard deviation for each feature and condition was presented. Moreover, a two-sample t-test to analyse significant mean differences at the group level was performed. T-tests were performed repeatedly across the 38 features, therefore, the Bonferroni correction was applied considering the significance level at $p$-value=0.05/38=0.0013.

Since results from these univariate analyses will be later contrasted with the classification approach, those features that the machine learning classifier considered as more relevant during the classification analysis were shown here in bold. Moreover, in order to later compare the distribution of relevant features between groups a multiple boxplot was computed in Figure 3.6. For each of the 7 features, their original values were standardized to z-scores. In each boxplot, outliers were plotted by means of individual dots.

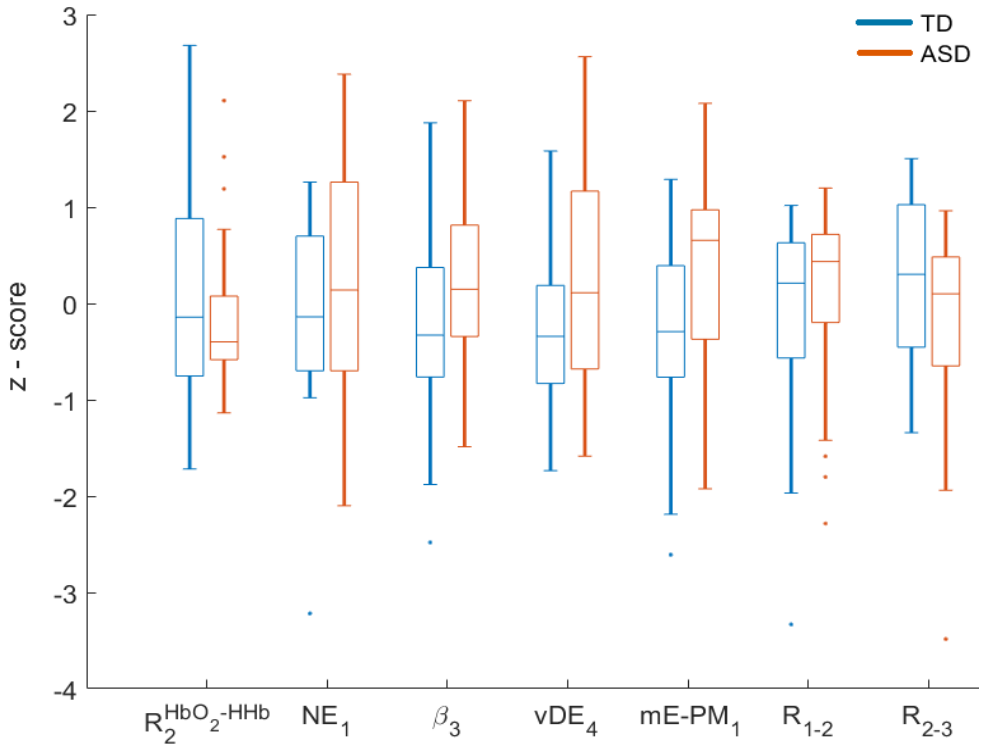



FIGURE 3.6: Multiple variables boxplot for the 7 fNIRS features that were selected in the classification analysis. Original features were standardized to z-scores, and the distribution of values was shown for each condition (TD and ASD) using a boxplots.

TABLE 3.2: Summary table for the 38 fNIRS features. The mean of each feature is shown together with the standard deviation for each condition. Mean values for each feature were compared across conditions using a two-sample t-test. The $p$-value resulting from this test was computed. $p$-value <0.0013 was considered a significant result. Features in bold were the ones later selected by the classification algorithm.

| N$^o$ | Feature ID | TD | ASD | $p$-value |
|---|---|---|---|---|
| 1 | $R_1^{HbO_2-HHb}$ | -0.34 ± 0.24 | -0.33 ± 0.18 | 0.761 |
| 2 | $\mathbf{R_2^{HbO_2-HHb}}$ | -0.31 ± 0.27 | -0.36 ± 0.18 | **0.385** |
| 3 | $R_3^{HbO_2-HHb}$ | -0.37 ± 0.21 | -0.37 ± 0.18 | 0.996 |
| 4 | $R_4^{HbO_2-HHb}$ | -0.40 ± 0.25 | -0.42 ± 0.13 | 0.724 |
| 5 | $\beta_1$ | -3.5 $10^{-5}$ ± 16.2 $10^{-5}$ | 6.3 $10^{-5}$ ± 15.5 $10^{-5}$ | 0.030 |
| 6 | $\beta_2$ | -2.7 $10^{-5}$ ± 13.6 $10^{-5}$ | 4.5 $10^{-5}$ ± 22.1 $10^{-5}$ | 0.164 |
| 7 | $\boldsymbol{\beta_3}$ | 2.8 $10^{-5}$ ± 15.1 $10^{-5}$ | 9.0 $10^{-5}$ ± 12.8 $10^{-5}$ | **0.113** |
| 8 | $\beta_4$ | 3.2 $10^{-5}$ ± 14.1 $10^{-5}$ | 3.6 $10^{-5}$ ± 10.2 $10^{-5}$ | 0.895 |
| 9 | $\mathbf{NE_1}$ | 19.52 ± 1.66 | 19.98 ± 1.91 | **0.357** |
| 10 | $NE_2$ | 19.60 ± 1.69 | 19.44 ± 2.16 | 0.776 |
| 11 | $NE_3$ | 20.19 ± 1.30 | 19.56 ± 1.76 | 0.150 |
| 12 | $NE_4$ | 19.85 ± 1.51 | 19.92 ± 1.57 | 0.869 |
| 13 | $mDE_1$ | 14.99 ± 2.57 | 14.80 ± 2.82 | 0.798 |
| 14 | $mDE_2$ | 14.70 ± 2.78 | 14.01 ± 3.76 | 0.454 |
| 15 | $mDE_3$ | 15.63 ± 1.86 | 15.13 ± 3.29 | 0.504 |
| 16 | $mDE_4$ | 14.10 ± 2.91 | 14.21 ± 2.36 | 0.883 |
| 17 | $vDE_1$ | 11.13 ± 2.80 | 12.13 ± 2.53 | 0.185 |
| 18 | $vDE_2$ | 11.87 ± 2.63 | 11.68 ± 2.40 | 0.789 |
| 19 | $vDE_3$ | 11.81 ± 2.43 | 11.85 ± 3.03 | 0.953 |
| 20 | $\mathbf{vDE_4}$ | 11.42± 1.84 | 12.5± 2.42 | **0.056** |
| 21 | $\mathbf{mE\text{-}PM_1}$ | -6.56± 4.67 | -3.6± 4.46 | **0.027** |
| 22 | $mE\text{-}PM_2$ | -4.35± 3.94 | -3.3± 5.88 | 0.471 |
| 23 | $vE\text{-}PM_1$ | 18.72± 4.95 | 20.3± 8.73 | 0.418 |
| 24 | $vE\text{-}PM_2$ | 18.29± 5.74 | 19.5± 7.09 | 0.477 |
| 25 | $mIE_1$ | 65.61± 4.69 | 65.2± 7.42 | 0.833 |
| 26 | $mIE_2$ | 67.07± 5.19 | 67.0± 7.21 | 1.000 |
| 27 | $mIE_3$ | 65.46± 4.59 | 66.1± 5.10 | 0.621 |
| 28 | $mIE_4$ | 64.25± 4.12 | 66.3± 4.37 | 0.088 |
| 29 | $vIE_1$ | 38.92± 10.28 | 40.4± 10.31 | 0.606 |
| 30 | $vIE_2$ | 40.00± 8.75 | 41.7± 10.04 | 0.500 |
| 31 | $vIE_3$ | 40.11± 7.30 | 42.3± 10.62 | 0.374 |
| 32 | $vIE_4$ | 40.56± 8.08 | 43.1± 10.49 | 0.320 |
| 33 | $\mathbf{R_{1-2}}$ | 0.65± 0.23 | 0.7± 0.20 | **0.415** |
| 34 | $R_{1-3}$ | 0.46± 0.18 | 0.3± 0.18 | 0.157 |
| 35 | $R_{1-4}$ | 0.39± 0.18 | 0.4± 0.21 | 0.391 |
| 36 | $\mathbf{R_{2-3}}$ | 0.44± 0.18 | 0.3± 0.21 | **0.053** |
| 37 | $R_{2-4}$ | 0.36± 0.19 | 0.4± 0.22 | 0.532 |
| 38 | $R_{3-4}$ | 0.41± 0.17 | 0.4± 0.18 | 0.804 |

## 3.2 Classification

The classification analyses were performed with first considering the behavioural and fNIRS features separately and comparing the performance with and without feature selection. Then, the most relevant features from each classification analysis were selected for an integrated multimodal approach.

### 3.2.1 Behavioural analysis

In Table 3.3, the main results from the classification approach with behavioural features were presented. The leave-one-out classification analysis using the LogitBoost algorithm was performed with first all the features and then performing a previous feature selection stage with the wrapper method. In Figure 3.7, the ROC curves obtained in both analyses were plotted together with the optimal operating point for each of the curves. During the feature selection stage, the wrapper method was required to select 4 features each time according to the results shown in Figure 2.11. As feature selection was applied in each of the 52 folds performed in the analysis, slightly different features might be selected by the algorithm in each fold. Therefore, in Table 3.4 the features that were selected in any of the folds were shown together with their frequency of selection. The ID of each of the behavioural features can be found in Section 2.3.1.

TABLE 3.3: Summary table with the results from the classification analysis applied to the behavioural feature set. Results with and without the feature selection stage were compared.

| Classification behavioural features | | |
|---|---|---|
| | All features | Feature selection |
| Accuracy (%) | 73 | 65 |
| Sensitivity (%) | 81 | 69 |
| Specificity (%) | 65 | 62 |
| AUC | 0.65 | 0.63 |

TABLE 3.4: Description of the behavioural features that the algorithm selected during the leave-one-out approach. Behavioural features that were not selected in any fold were not included in this table.

| Features selected | Frequency (% of folds) |
|---|---|
| **$SLP_1$** | **100** |
| **vPM** | **100** |
| **ACC** | **100** |
| **mIBV** | **86.5** |
| vIBV | 13.5 |

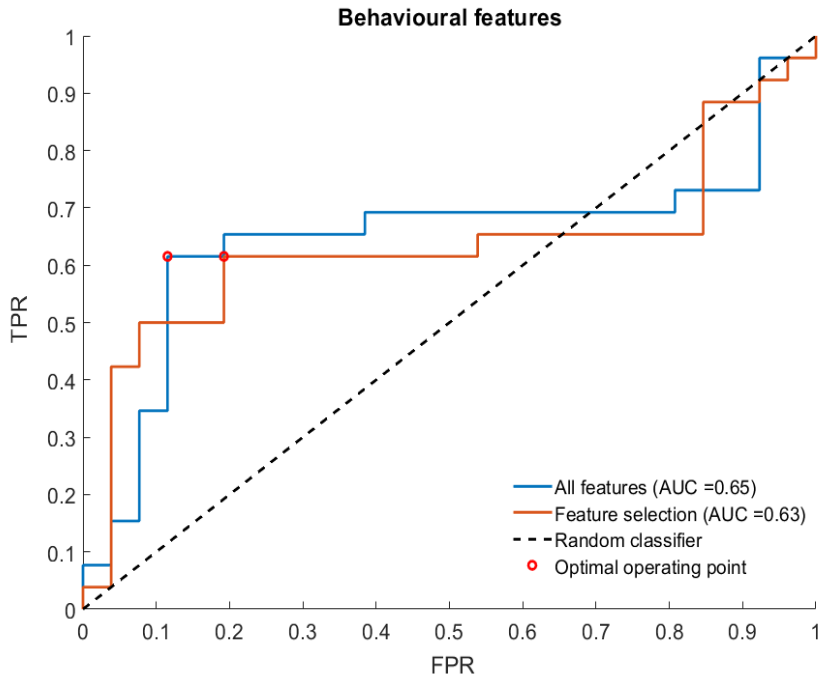



FIGURE 3.7: Receiving operating characteristic (ROC) curve when the LogitBoost algorithm discriminated between TD and ASD subjects on the basis of their behavioural features. True positive rate (TPR) was computed against the false positive rate (FPR) and the optimal operating point was plotted. The area under the ROC curve (AUC) was shown for both analyses.

### 3.2.2 fNIRS analysis

The same classification procedure than in the previous section was then applied to the 38 fNIRS feature set. The performance of the algorithm on discriminating between TD and ASD was summarized in Table 3.5. Comparison with and without performing feature selection is shown in the table. ROC curves from both classification analyses were plotted in Figure 3.8. Finally, during the classification analysis including feature selection the algorithm was required to select 7 features each time, as justified in Figure 2.11. Hence, features that were selected at least in one of the 52 folds were presented in Figure 3.6 together with the frequency in which the algorithm selected them (proportion of selection out of the 52 folds). The ID of each of the fNIRS features can be found in Section 2.3.2.

TABLE 3.5: Summary table with the results from the classification analysis applied to the fNIRS feature set. Results with and without the feature selection stage were compared.

| Classification fNIRS features | | |
|---|---|---|
| | All features | Feature selection |
| Accuracy (%) | 69 | 60 |
| Sensitivity (%) | 77 | 65 |
| Specificity (%) | 62 | 54 |
| AUC | 0.67 | 0.65 |

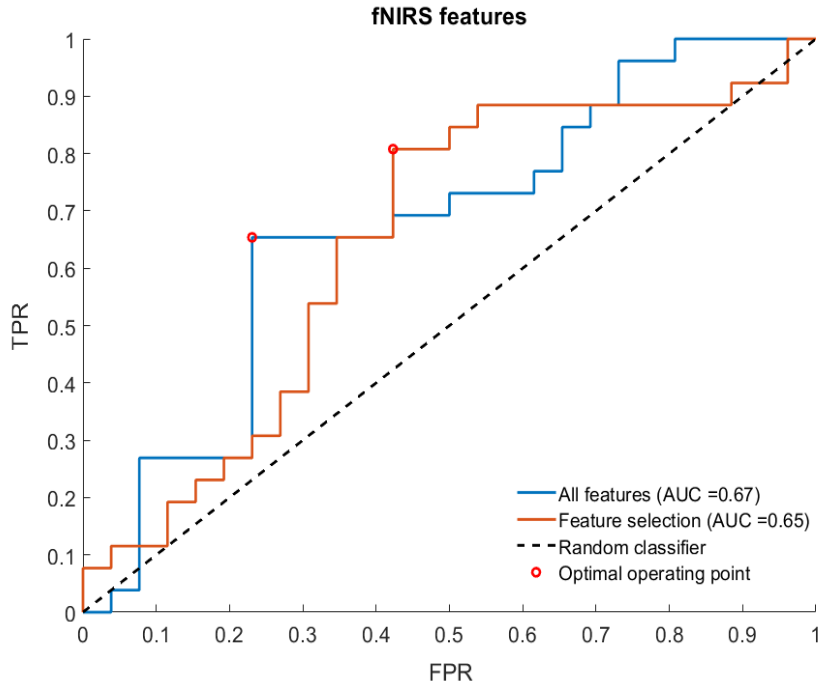



FIGURE 3.8: Receiving operating characteristic (ROC) curves for the classification analysis with fNIRS features. True positive rate (TPR) was computed against the false positive rate (FPR) and the optimal operating point was plotted. The area under the ROC curve (AUC) was shown for both analyses.

TABLE 3.6: Description of the fNIRS features that the algorithm selected during the leave-one-out approach. Seven features were selected in each of the 52-folds. The 7 features that appeared more often are shown in bold. fNIRS features that were not selected in any fold were not included in this table.

| Features selected | Frequency (% of folds) | Features selected | Frequency (% of folds) |
|---|---|---|---|
| **mE-PM$_1$** | **100** | R$_{1-3}$ | 5.77 |
| **R$_{1-2}$** | **84.62** | vDE$_1$ | 3.85 |
| **R$_2^{HbO_2-HHb}$** | **84.62** | mDE$_3$ | 3.85 |
| **vDE$_4$** | **82.69** | mDE$_1$ | 3.85 |
| **R$_{2-3}$** | **78.85** | $\beta_2$ | 3.85 |
| **$\beta_3$** | **71.15** | NE$_2$ | 3.85 |
| **NE$_1$** | **65.38** | R$_{2-4}$ | 1.92 |
| $\beta_1$ | 25 | vIE$_1$ | 1.92 |
| NE$_3$ | 23.08 | vE-PM$_1$ | 1.92 |
| mIE$_4$ | 15.38 | mE-PM$_2$ | 1.92 |
| R$_{1-4}$ | 11.54 | mDE$_4$ | 1.92 |
| R$_1^{HbO_2-HHb}$ | 11.54 | $\beta_4$ | 1.92 |
| vE-PM$_2$ | 7.69 | NE$_4$ | 1.92 |

### 3.2.3 Integrated analysis

An integrated analysis including behavioural and fNIRS features was performed. Due to the fact that for the fNIRS analysis the algorithm did not show high consistency in the feature selection stage, not all the features were included in this analysis. Instead, from the behavioural set, only the 4 features that were selected more often were included in this occasion (Table 3.4) together with the top 7 fNIRS features (Table 3.6). Hence, the 11 features included in this analysis were: $SLP_1$, vPM, ACC, mIBV for the behavioural set, and mE-$PM_1$, $R_{1-2}$, $R_2^{HbO_2-HHb}$, $vDE_4$, $R_{2-3}$, $\beta_3$ and $NE_1$ for the fNIRS set. The results obtained during this classification analysis were presented in Table 3.7, and the ROC curve was illustrated in Figure 3.9 along with the optimal operating point.

Table 3.7: Classification results from the integrated analysis.

| **Classification integrated analysis** | |
|---|---|
| Accuracy (%) | 79 |
| Sensitivity (%) | 81 |
| Specificity (%) | 77 |
| AUC | 0.8 |

Finally, it is important to note that selecting the best features from the behavioural and fNIRS analyses and joining them together to perform a single classification approach is likely to induce bias. As previously mentioned in the methods section, feature selection should only be performed on the training set, and it was performed in this way for the behavioural and fNIRS analysis. However, for this analysis, including those features that were selected more often in each of the previous experiments can also contain bias, as results from the behavioural and fNIRS analyses were summarized for the whole set. Therefore, in order to detect whether the process contained bias and generated a null distribution, a permutation test was performed. This test consisted on repeating the whole process 10.000 times until obtaining the accuracy from the integrated analysis. However, at each time, the labels of the participants (TD/ASD) were randomly permuted. After this whole process was performed, a histogram with the accuracies obtained in each of the 10.000 permutation tests was plotted along with the classification performance obtained with the real labels. In Figure 3.10, the results from the permutation test can be observed, and the accuracy obtained with the real class labels was plotted with a black arrow. The histogram from Figure 3.10 can be treated as the null hypothesis distribution of accuracies of the procedure. Hence, the statistical significance of the accuracy obtained with the real labels can be obtained by computing the proportion of permutation test accuracies that were higher than the one from the real classification test (79 %). The p-value obtained with this procedure was 0.014, which was the proportion of accuracies higher than 79 % over the 10.000 trials performed.

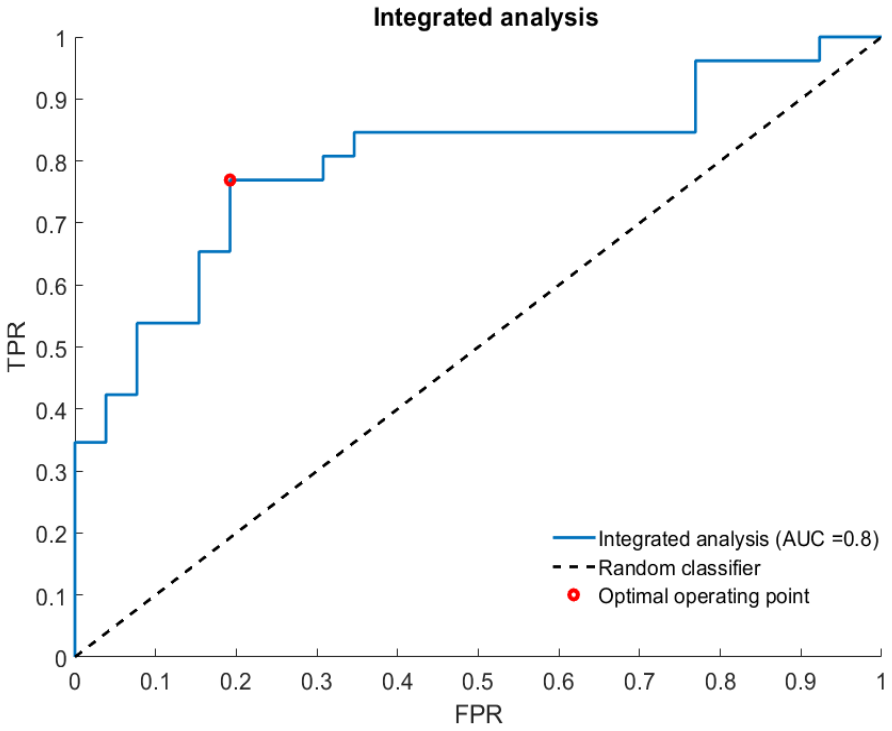



FIGURE 3.9: Receiving operating characteristic (ROC) curve when the LogitBoost algorithm discriminated between TD and ASD subjects with the integrated features. True positive rate (TPR) was computed against the false positive rate (FPR) and the optimal operating point was plotted. The area under the ROC curve (AUC) was shown.

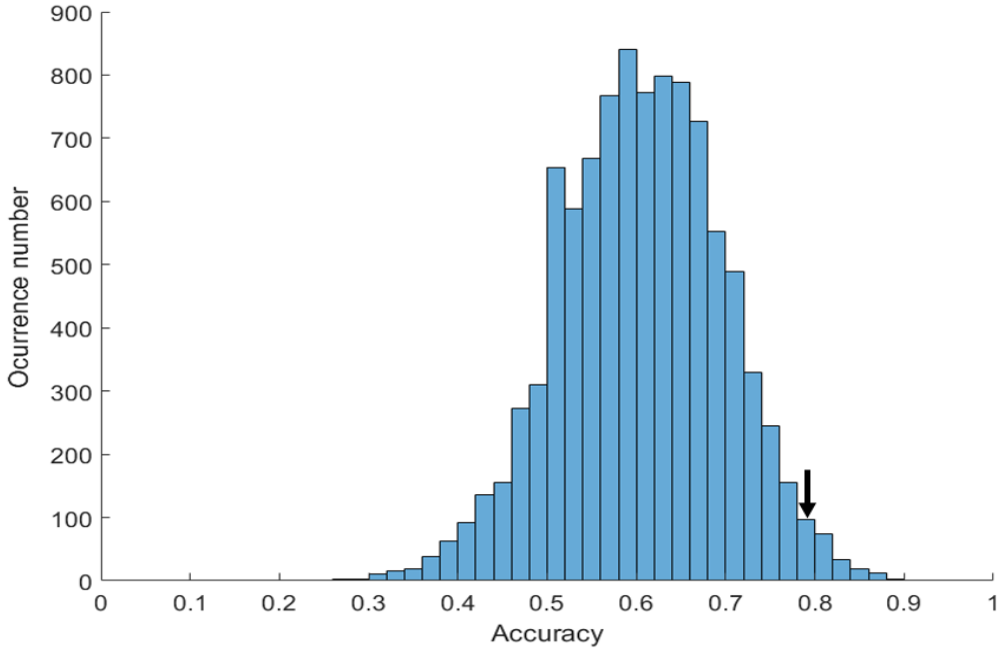



FIGURE 3.10: Histogram showing the distribution of accuracies after repeating the same procedure 10.000 rimes permuting the labels of each subject. The histogram represents the null hypothesis and the black arrow the accuracy obtained with the real class labels. The proportion of accuracies larger than the arrow was 0.014 (p-value).

# Discussion

In this study, machine learning methods were applied to discriminate between 26 TD and 26 high-functioning ASD subjects on the basis of 11 and 38 features extracted from their behavioural and neuroimaging data, respectively. Visualization and summary statistics techniques were first used to observe tendencies, detect characteristic patterns and provide support for the interpretation of results from the classification analyses.

In Figure 3.1, response times (RTs) to the ongoing (OG) tasks were sequentially plotted for each participant and condition throughout the experiment. In most of the cases, sequential patterns on RTs were observed, but most importantly, changes in trend, frequency and variability were noticed between blocks (with different social scenarios) exhibiting the effect of social conditions on the OG task. Moreover, by comparing Figure 3.1a with Figure 3.1b, a particularly high degree of variability on RTs was detected for some ASD participants, between and within blocks. In fact, Figure 3.3 exhibits 8 ASD subjects with an atypically high intra-block variance. Interestingly, a different linear trend was observed in this figure when comparing mean intra block-variance (mIBV) and variance of intra block-variances (vIBV) between groups. It is detected how the vIBV increases more rapidly as the mIBV raises in the TD group, suggesting more changes in the behaviour of TD participants when social conditions were altered (blocks). In addition, time intervals of the PM task were plotted in Figure 3.4. From those graphs, a high degree of heterogeneity was detected between ASD participants. Whereas some subjects exhibited a very consistent and precise pattern throughout the experiment (e.g. 180, 313, 314), others presented very variable PM intervals (e.g. 351, 548, 558). It is noteworthy that some ASD participants that exhibited an atypically high RT variance, presented a standard behaviour related to the PM task, and vice-versa. This observation suggests that multiple aspects of a certain domain need to be considered jointly to better characterize the ASD group. The distribution of mean accuracies in the OG task was compared for each condition in the boxplot from Figure 3.5. Although both groups presented a very high accuracy, it is also observed how the lower whisker from the ASD boxplot includes slightly lower accuracies than the TD group, showing the atypical performance of some of the ASD participants. Finally, from the summary statistics tables of behavioural and fNIRS features (Table 3.1 and 3.2), no significant mean differences were detected at the group level. All together, these results exhibited the heterogeneous and multifactorial pattern of ASD deficits in this study, which is often reported in the literature [8].

Before applying the classification procedure, the optimal number of features that the wrapper algorithm included in each iteration of the classification analysis was decided. From this study, important results were detected. However, note that these results were presented in the methods section to facilitate the comprehension of subsequent methods. In Figure 2.11, it was shown that a number of 4 and 7 features were the optimal subset sizes that minimized misclassification rate (MCR). For computing the MCR of a subset size of 1 feature, the SFS algorithm performed an exhaustive search in which all the individual features were analysed separately, and their MCR was obtained. It is important to note that in this analysis, the introduction of more than one feature in the candidate set always reduced MCR, in both behavioural and fNIRS sets. This result remarked the relevance of multivariate techniques for discriminating ASD subjects, which exhibited better performance than univariate classification.

## 4.1 Behavioural discriminance

Classification analyses were initially performed including all the extracted behavioural features and later using a previous feature selection stage. The performance of both analyses was summarized in Table 3.3. When considering all the behavioural features, the algorithm achieved an overall accuracy of 73% and an AUC from the ROC curve of 0.65. It is important to note that the algorithm exhibited higher sensitivity (81%) than specificity (65%), being able to better identify ASD than TD subjects. This imbalance is probably caused by the common distribution that both groups exhibit, in which a certain proportion of ASD participants show an atypical behaviour and are clearly differentiated by the algorithm, whereas those TD and ASD subjects that behave similarly are all classified with lower accuracy.

When including the feature selection stage, accuracy, sensitivity, specificity and AUC all slightly decreased. However, as the AUCs and ROC curves from both analyses were similar (see Figure 3.7), their discriminance can be considered to have yielded similarly (as ROC analyses are normally better indicators of the relation between accurate predictions and the posterior probabilities behind each prediction). This lack of improvement when feature selection was applied is likely to occur because the classifier was able to identify the relevant features over the whole set. Hence, it is likely that the classifiers trained in each experiment encoded for similar information, and when all features were considered, the LogitBoost algorithm applied more weights to those nodes splitting the 4 relevant features.

In Table 3.4, features that were selected in any of the 52 folds were presented together with their frequency of selection. It is observed how in almost all folds the same features

were selected by the algorithm. Therefore, it is likely that these features were the ones encoding for discriminant information. The 4 features selected more often included: the slope of the RT during the first part of the experiment ($SLP_1$), the variance in the PM hits (vPM), the mean accuracy in the OG task (ACC) and the mean intra-block variance (mIBV). It is observed how in 7 of the 52 folds, vIBV was selected instead of mIBV. This fact is probably caused by the high correlation between these two, as discussed in Figure 3.3. From these results, it is detected how measures of dispersion and intra-individual variability proved to be more discriminant at the single subject level than measures of central tendency (like mRT or mPM). Therefore, behavioural functional biomarkers for ASD diagnosis are more likely to be measures of dispersion than mean values.

Finally, it is interesting to observe that the features that proved to be more discriminant by the algorithm were not always the ones exhibiting more differences at the group level analysis. In Table 3.1, it is observed how features with relatively high p-values (e.g. $SLP_1$), were selected instead of RMSE, which showed the lowest p-value at the group level. This result displays the limitations of univariate methods when identifying discriminant biomarkers, and it shows that features that individually show little group differences can have a discriminant role in combination with others.

## 4.2 fNIRS discriminance

Analogous to the behavioural analyses, classification performance was evaluated for the fNIRS features with and without the feature selection stage. The main results from these analyses were shown in Table 3.5. For both cases, higher sensitivity than specificity on the discrimination approach was detected.

When all features were included, the accuracy and AUC obtained were 69% and 0.67, respectively. On the other hand, for the analysis including feature selection, the accuracy obtained was 60% and the AUC 0.65. Hence, feature selection also resulted in slightly worse performance for the classification task. However, both ROC curves had similar shapes and magnitudes (Figure 3.8), and therefore similar AUCs were obtained. Despite higher degrees of sensitivity were obtained with and without the feature selection stage, the ROC curves for the fNIRS features exhibited a more balanced shape than during the behavioural analysis, indicating more resilience to class imbalance.

In Table 3.6, the features that were selected during any of the 52-fold analyses were presented with their frequency of selection. A high degree of inconsistency on the features selected was observed during this analysis. Inconsistent feature selection is likely to reduce the performance of the algorithm, and it might be the cause for the lower

accuracy observed in this analysis (compared to the accuracy obtained when using all features). One of the main reasons behind this inconsistent pattern might be the number of variables used and the relatively small number of subjects. Given the small number of observations, using 38 features as an input for the classification algorithm might significantly worsen its performance on detecting consistent patterns, as many parameters need to be optimized and not enough observations might be provided to perform this learning process. On the other hand, it could also be an indicator of little discriminant patterns on the feature space.

Nevertheless, a group of 7 features did seem to appear much more often than the others (Table 3.6). Only one of the fNIRS features appeared in 100% of the folds, which was mE-$PM_1$. In addition, although no significant differences were detected, this was the feature with the lowest p-value on the group mean statistical test (see Table 3.2). This is particularly interesting, as mE-$PM_{1-2}$ and vE-$PM_{1-2}$ were the only types of features that integrated behavioural and fNIRS information into one single value. mE-PM computed the difference between the time point (in seconds) in which a functional neural event was detected minus the time point of the nearest PM hit (behavioural event). As the prospective memory (PM) task involved forming mental intentions to perform future actions, it would be expected that functional neural activation occurred before the behavioural action (PM hit). From Table 3.2, it was seen how the mean value of mE-$PM_1$ for the TD group was -6.56 seconds whereas for the ASD was -3.6 seconds. In terms of tasks related to the PM, the more ability to anticipate and imagine the future scenario, the more distant the functional activation will be from the behaviour itself [39]. Therefore, results from this feature suggest a greater relation between neural activation and the behavioural PM task for the TD group, which at the same time resulted to be the feature with more discriminative power (Table 3.6). Thus, one of the most notable differences between ASD and TD exhibited in this analysis might be the time distance between functional activation and behavioural PM task. This result provided evidence from the present methodology to detect discriminant features that could be later used to understand atypical brain-behaviour interactions in ASD.

Both regional correlations, $R_{1-2}$ and $R_{2-3}$ also seemed to be particularly discriminant. These results were consistent with atypical synchrony patterns in these regions previously detected in ASD. Atypical co-activation between these regions of the PFC in ASD subjects was reported in previous studies [24]. Another feature that appeared with high frequency was $vDE_4$. As observed in Figure 3.6, higher degree of variation in the duration of functional events in ASD subjects in the right area of the PFC was detected. In addition, the fact that $\beta_3$ appeared in the feature set indicated differences between the groups when analysing the effect of the reward. Lower $\beta_3$ values for the TD group

(Figure 3.6) suggested that the fNIRS signal in region 3 exhibited a higher degree of contrast between those scenarios in which participants were earning money for themselves or for others. This result reinforces the hypothesis that ASD subjects might react less mindfully to the effects of social manipulations. The number of functional events in the central R1 region ($NE_1$) also provided significant discriminant capacity. A much higher degree of variability was detected in this feature for the ASD group. As R1 is involved in mental processes related to the PM, this result suggests less consistency in the ASD group during PM tasks, which corresponds to the results obtained in the behavioural analysis. Moreover, together with other results, this fact indicates that the relation between specific behavioural and fNIRS features is clear, and therefore, using features that consider both modalities is likely to synthesize and provide important information with a lower number of variables.

Overall, these results displayed the discriminant capacity of certain features that did not show significant differences at the group level when univariate t-tests were performed. This fact elucidates the unique capacity of this methodology to study potential biomarkers for ASD. On the other hand, the small number of observations and highly dimensional fNIRS set limited the capacity of the algorithm to discriminate between both groups.

## 4.3 Integrated classification

Finally, the best behavioural and fNIRS features were used together to perform an integrated classification analysis. Table 3.7 summarizes the results from this analysis. The approach exhibited a performance of 79% accuracy, 81% sensitivity, 77% specificity and 0.8 AUC. These results might seem to highly outperform any of the behavioural or fNIRS analyses alone. However, it is important to note that this procedure might contain bias, as the features selected resulted from the 52-fold analysis on the behavioural and fNIRS sets, and therefore selection was not applied on a training set. In order to analyse bias on feature selection and consider the significance of the performance obtained, a permutation test was performed as described in Section 3.2.3. From the histogram obtained in Figure 3.10, it can be observed how the distribution of accuracies during the permutation test was not centred at 50%, which would be the value expected if the procedure did not contain bias. Instead, the null distribution was centred around 60% accuracy, indicating a procedural bias of $\approx 10\%$.

By obtaining the null distribution (Figure 3.10), the significance of the accuracy achieved with the real class labels could be analysed, and a p-value of 0.014 was obtained. This result indicated that significant discriminance was achieved. With 98.6% confidence, it

can be considered that the accuracy obtained in the classification analysis (79%) was not likely to be caused by a random effect. However, it was also observed that this accuracy contained a procedural bias around 10%. This fact suggests that if the experiment was repeated, and the same features from this analysis were used, significant discriminance is likely to be obtained, but the accuracy in the classification would be more likely to be closer to 69 than 79%.

These results showed that the performance obtained with the integrated multi-modal analysis did not improve compared to the behavioural or fNIRS classification separately, since they both obtained similar accuracies. Despite this approach aimed to detect complementary information of ASD deficits contained in different domains, the results obtained suggested that the behavioural and neuroimaging analyses were both encoding for similar information. The fact that correlated results were obtained in the behavioural and fNIRS analyses separately indicates that some features from both sets might contain the same information about ASD deficits. This occurence would be likely to limit the capacity of the algorithm to find independent discriminant variables across domains. Hence, future multimodal platforms accounting for complementary information about ASD deficits might require to consider inter-modality correlations, enabling classification algorithms to better find those patterns that differ across domains.

# Conclusions

The machine learning approach implemented in this study demonstrated that multivariate analyses considering multiple aspects of one modality can better characterize ASD subjects than using single variables. The analyses performed exhibited the capacity of this methodology to study the most discriminant features of each modality, identifying potential biomarkers for ASD diagnosis that were not detected by conventional univariate tests. In addition, the methods developed provided a null distribution for the analysis, and significant discriminance between TD and ASD subjects was achieved. On the other hand, the integration of features from behavioural and neuroimaging data did not show an improvement on the classification, which suggested that the features extracted in each modality encoded for similar information of ASD deficits. Future studies aiming to develop multi-modal diagnostic platforms for ASD would probably benefit from considering dependencies between features across domains, maximizing the information provided by each modality and providing a broader view into the whole spectrum.

# Bibliography


[1] Sam J. Gilbert, Julia D.I. Meuwese, Karren J. Towgood, Christopher D. Frith, and Paul W. Burgess. Abnormal functional specialization within medial prefrontal cortex in high-functioning autism: A multi-voxel similarity analysis. *Brain*, 2009. ISSN 00068950. doi: 10.1093/brain/awn365.

[2] Francesca Happé and Angelica Ronald. The 'fractionable autism triad': A review of evidence from behavioural, genetic, cognitive and neural research, 2008. ISSN 10407308.

[3] Daniel H. Geschwind. Genetics of autism spectrum disorders, 2011. ISSN 13646613.

[4] Marco Catani et.al. Frontal networks in adults with autism spectrum disorder. *Brain*, 139(2):616–630, 2016. ISSN 14602156. doi: 10.1093/brain/awv351.

[5] Paolo Brambilla, Antonio Hardan, Stefania Ucelli Di Nemi, Jorge Perez, Jair C. Soares, and Francesco Barale. Brain anatomy and development in autism: Review of structural MRI studies, 2003. ISSN 03619230.

[6] Sam J. Gilbert, Geoffrey Bird, Rachel Brindley, Christopher D. Frith, and Paul W. Burgess. Atypical recruitment of medial prefrontal cortex in autism spectrum disorders: An fMRI study of two executive function tasks. *Neuropsychologia*, 2008. ISSN 00283932. doi: 10.1016/j.neuropsychologia.2008.03.025.

[7] Karren J. Towgood, Julia D.I. Meuwese, Sam J. Gilbert, Martha S. Turner, and Paul W. Burgess. Advantages of the multiple case series approach to the study of cognitive deficits in autism spectrum disorder. *Neuropsychologia*, 47(13):2981–2988, 2009. ISSN 00283932. doi: 10.1016/j.neuropsychologia.2009.06.028. URL `http://dx.doi.org/10.1016/j.neuropsychologia.2009.06.028`.

[8] Christine Ecker and Declan Murphy. Neuroimaging in autism-from basic science to translational research, 2014. ISSN 17594758.

[9] Paola Pinti, Clarisse Aichelburg, Sam Gilbert, Antonia Hamilton, Paul Burgess, and Ilias Tachtsidis. A review of functional Near-Infrared Spectroscopy measurements in naturalistic environments. pages 1–63.

[10] Yusuke Moriguchi and Kazuo Hiraki. Prefrontal cortex and executive function in young children: a review of NIRS studies. *Frontiers in Human Neuroscience*, 2013. ISSN 1662-5161. doi: 10.3389/fnhum.2013.00867.

[11] Paola Pinti, Clarisse Aichelburg, Sam Gilbert, Antonia Hamilton, Joy Hirsch, Paul Burgess, and Ilias Tachtsidis. A Review on the Use of Wearable Functional Near-Infrared Spectroscopy in Naturalistic Environments. *Japanese Psychological Research*, 2018. ISSN 00215368. doi: 10.1111/jpr.12206. URL `http://doi.wiley.com/10.1111/jpr.12206`.

[12] Paul W. Burgess and Donald T. Stuss. Fifty years of prefrontal cortex research: Impact on assessment, 2017. ISSN 14697661.

[13] Felix Scholkmann, Stefan Kleiser, Andreas Jaakko Metz, Raphael Zimmermann, Juan Mata Pavia, Ursula Wolf, and Martin Wolf. A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology, 2014. ISSN 10538119.

[14] Felix Scholkmann and Martin Wolf. General equation for the differential pathlength factor of the frontal human head depending on wavelength and age. *Journal of Biomedical Optics*, 2013. ISSN 1083-3668. doi: 10.1117/1.JBO.18.10.105004.

[15] Ferran Gonzalez Hernandez. Multivariate approach to characterize Autism Spectrum Disorder. *CoMPLEX MRes ,3rd mini-project*, 2018.

[16] Mohammad R. Arbabshirani, Sergey Plis, Jing Sui, and Vince D. Calhoun. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 2017. ISSN 10959572. doi: 10.1016/j.neuroimage.2016.02.079.

[17] Honghui Yang, Jingyu Liu, Jing Sui, Godfrey Pearlson, and Vince D. Calhoun. A Hybrid Machine Learning Method for Fusing fMRI and Genetic Data: Combining both Improves Classification of Schizophrenia. *Frontiers in Human Neuroscience*, 2010. ISSN 1662-5161. doi: 10.3389/fnhum.2010.00192.

[18] Vince D. Calhoun and Jing Sui. Multimodal Fusion of Brain Imaging Data: A Key to Finding the Missing Link(s) in Complex Mental Illness, 2016. ISSN 24519022.

[19] Nader Karamzadeh, Franck Amyot, Kimbra Kenney, Afrouz Anderson, Fatima Chowdhry, Hadis Dashtestani, Eric M. Wassermann, Victor Chernomordik, Claude Boccara, Edward Wegman, Ramon Diaz-Arrastia, and Amir H. Gandjbakhche. A machine learning approach to identify functional biomarkers in human prefrontal cortex for individuals with traumatic brain injury using functional near-infrared spectroscopy. *Brain and Behavior*, 6(11):1–14, 2016. ISSN 21623279. doi: 10.1002/brb3.541.

[20] Hirokazu Atsumori. Noninvasive imaging of prefrontal activation during attention-demanding tasks performed while walking using a wearable optical topography system. *Journal of Biomedical Optics*, 2010. ISSN 1083-3668. doi: 10.1117/1.3462996.

[21] Theodore J Huppert, Solomon G Diamond, Maria Angela Franceschini, and David A Boas. HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain. *Applied optics*, 2009. ISSN 08966273. doi: 10.1016/j.drugalcdep.2008.02.002.A.

[22] Paola Pinti, Arcangelo Merla, Clarisse Aichelburg, Frida Lind, Sarah Power, Elizabeth Swingler, Antonia Hamilton, Sam Gilbert, Paul W. Burgess, and Ilias Tachtsidis. A novel GLM-based method for the Automatic IDentification of functional Events (AIDE) in fNIRS data recorded in naturalistic environments. *NeuroImage*, 155(December 2016):291–304, 2017. ISSN 10959572. doi: 10.1016/j.neuroimage.2017.05.001. URL `http://dx.doi.org/10.1016/j.neuroimage.2017.05.001`.

[23] Paul W. Burgess, Sophie K. Scott, and Christopher D. Frith. The role of the rostral frontal cortex (area 10) in prospective memory: A lateral versus medial dissociation. *Neuropsychologia*, 2003. ISSN 00283932. doi: 10.1016/S0028-3932(02)00327-5.

[24] Sam J. Gilbert, Gil Gonen-Yaacovi, Roland G. Benoit, Emmanuelle Volle, and Paul W. Burgess. Distinct functional connectivity associated with lateral versus medial rostral prefrontal cortex: A meta-analysis. *NeuroImage*, 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.07.032.

[25] Paul W. Burgess, Gil Gonen-Yaacovi, and Emmanuelle Volle. Functional neuroimaging studies of prospective memory: What have we learnt so far? *Neuropsychologia*, 49(8):2246–2257, 2011. ISSN 00283932. doi: 10.1016/j.neuropsychologia.2011.02.014.

[26] Marco Catani, Flavio Dell'Acqua, Francesco Vergani, Farah Malik, Harry Hodge, Prasun Roy, Romain Valabregue, and Michel Thiebaut de Schotten. Short frontal lobe connections of the human brain. *Cortex*, 48(2):273–291, 2012. ISSN 00109452. doi: 10.1016/j.cortex.2011.12.001.

[27] K. J. Friston, a. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 1995. ISSN 10659471. doi: 10.1002/hbm.460020402.

[28] Xu Cui, Signe Bray, and Allan L. Reiss. Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *NeuroImage*, 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.11.050.

[29] Yue Gu, Shuo Miao, Junxia Han, Zhenhu Liang, Gaoxiang Ouyang, Jian Yang, and Xiaoli Li. Identifying ADHD children using hemodynamic responses during a working memory task measured by functional near-infrared spectroscopy. *Journal of neural engineering*, 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aa9ee9. URL `http://iopscience.iop.org/article/10.1088/1741-2552/aa9ee9{%}0Ahttp://www.ncbi.nlm.nih.gov/pubmed/29199636`.

[30] Ping Li. Robust logitboost and adaptive base class (abc) logitboost. *arXiv preprint arXiv:1203.3491*, 2012.

[31] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting, 2000. ISSN 00905364.

[32] Yu Dong Cai, Kai Y. Feng, Wen Cong Lu, and Kuo Chen Chou. Using LogitBoost classifier to predict protein structural classes. *Journal of Theoretical Biology*, 2006. ISSN 00225193. doi: 10.1016/j.jtbi.2005.05.034.

[33] Greg Ridgeway. The state of boosting. *Computing Science and Statistics*, 1999. doi: citeulike-article-id:7678637.

[34] B P Roe, H.-J. Yang, and J Zhu. Boosted Decision Trees, A Powerful Event Classifier. *Statistical Problems in Particle Physics, Astrophysics and Cosmology*, page 139, 2006. doi: 10.1142/9781860948985_0029. URL `http://adsabs.harvard.edu/abs/2006sppp.conf..139R`.

[35] Tom M. Mitchell. Decision Tree Learning, 1997.

[36] J Bergstra, Daniel L K Yamins, and D D Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *ICML*, 2013. ISSN 1938-7228.

[37] Mathworks. MATLAB - Mathworks - MATLAB & Simulink, 2016.

[38] Osamu Komori and Shinto Eguchi. A boosting method for maximizing the partial area under the ROC curve. *BMC Bioinformatics*, 2010. ISSN 14712105. doi: 10.1186/1471-2105-11-314.

[39] Jiro Okuda, Toshikatsu Fujii, Hiroya Ohtake, Takashi Tsukiura, Atsushi Yamadori, Christopher D. Frith, and Paul W. Burgess. Differential involvement of regions of rostral prefrontal cortex (Brodmann area 10) in time- and event-based prospective memory. *International Journal of Psychophysiology*, 2007. ISSN 01678760. doi: 10.1016/j.ijpsycho.2006.09.009.