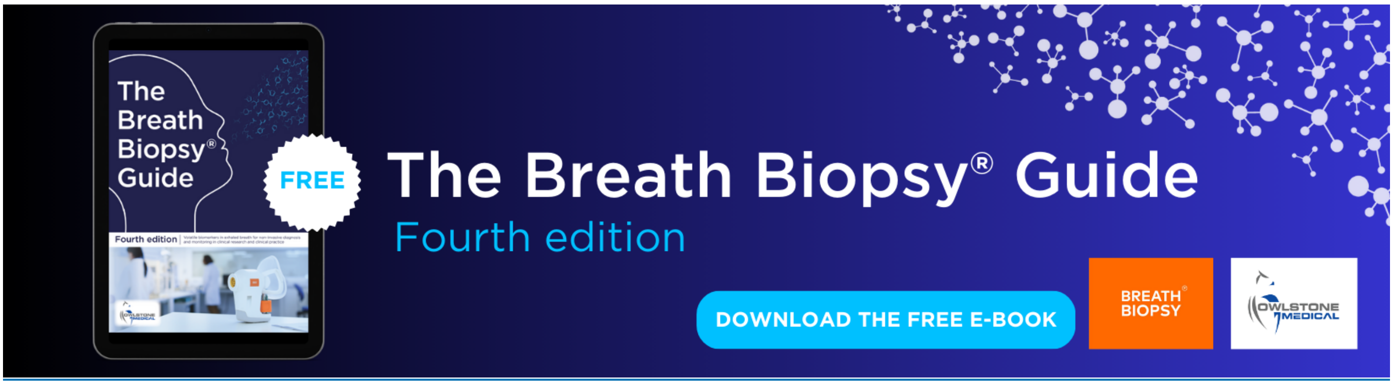
Journal of Neural Engineering

PAPER • OPEN ACCESS

Identifying ADHD boys by very-low frequency prefrontal fNIRS fluctuations during a rhythmic mental arithmetic task

To cite this article: Sergio Ortuño-Miró *et al* 2023 *J. Neural Eng.* **20** 036018



Journal of Neural Engineering

OPEN ACCESS

RECEIVED
28 September 2022

REVISED
13 December 2022

ACCEPTED FOR PUBLICATION
20 December 2022

PUBLISHED
23 May 2023

PAPER

Identifying ADHD boys by very-low frequency prefrontal fNIRS fluctuations during a rhythmic mental arithmetic task

Sergio Ortuño-Miró[1], Sergio Molina-Rodríguez[2], Carlos Belmonte[2] and Joaquín Ibañez-Ballesteros[1,*]

[1] Department of physiology, Miguel Hernandez University, San Joan d´Alacant, Alicante, Spain
[2] Institute of Neurosciences, Spanish National Research Council- Miguel Hernandez University, San Joan d´Alacant, Alicante, Spain
[*] Author to whom any correspondence should be addressed.

E-mail: charly.joa@umh.es

Keywords: attention-deficit/hyperactivity disorder, functional near-infrared spectroscopy, machine learning, classification, mental arithmetic, logistic regression, linear discriminant analysis

Abstract

Objective. Computer-aided diagnosis of attention-deficit/hyperactivity disorder (ADHD) aims to provide useful adjunctive indicators to support more accurate and cost-effective clinical decisions. Deep- and machine-learning (ML) techniques are increasingly used to identify neuroimaging-based features for objective assessment of ADHD. Despite promising results in diagnostic prediction, substantial barriers still hamper the translation of the research into daily clinic. Few studies have focused on functional near-infrared spectroscopy (fNIRS) data to discriminate ADHD condition at the individual level. This work aims to develop an fNIRS-based methodological approach for effective identification of ADHD boys via technically feasible and explainable methods. *Approach.* fNIRS signals recorded from superficial and deep tissue layers of the forehead were collected from 15 clinically referred ADHD boys (average age 11.9 years) and 15 non-ADHD controls during the execution of a rhythmic mental arithmetic task. Synchronization measures in the time-frequency plane were computed to find frequency-specific oscillatory patterns maximally representative of the ADHD or control group. Time series distance-based features were fed into four popular ML linear models (support vector machine, logistic regression (LR), discriminant analysis and naïve Bayes) for binary classification. A 'sequential forward floating selection' wrapper algorithm was adapted to pick out the most discriminative features. Classifiers performance was evaluated through five-fold and leave-one-out cross-validation (CV) and statistical significance by non-parametric resampling procedures. *Main results.* LR and linear discriminant analysis achieved accuracy, sensitivity and specificity scores of near 100% ($p < .001$) for both CV schemes when trained with only three key wrapper-selected features, arising from surface and deep oscillatory components of very low frequency. *Significance.* We provide preliminary evidence that very-low frequency fNIRS fluctuations induced/modulated by a rhythmic mental task accurately differentiate ADHD boys from non-ADHD controls, outperforming other similar studies. The proposed approach holds promise for finding functional biomarkers reliable and interpretable enough to inform clinical practice.

1. Introduction

Attention-deficit/hyperactivity disorder (ADHD) is recognized as a highly prevalent neurodevelopmental disorder in school-age children worldwide, often persisting into adolescence and adulthood, and frequently overlapped with other psychiatric comorbidities [1–3]. People with ADHD exhibit three core behavioral symptoms, inattention, hyperactivity and impulsivity, although each displayed to a varying degree [4]. ADHD is a highly heterogeneous impairing condition, extensively researched over many years, that is probably the best known childhood-onset disorder [5]. Based on the last 10–20 years of scientific evidence, updated information on ADHD was recently summarized by the 'World Federation of

ADHD' as an international consensus statement [6] and reviewed by Posner *et al* in [7], covering epidemiology, etiology, pathophysiology, diagnosis and treatment. It is currently accepted that ADHD is a complex, heterogeneous disorder, in which different expressions of impairment along with variable trajectories must be recognized in order to adopt personalized approaches that best target an individual. This is of crucial importance because, even despite serious distress/impairments, many patients lead rewarding and productive lives when properly managed.

Diagnosis of ADHD is event today based mainly on clinical signs and symptoms that require a detailed evaluation by an expert clinician through interviews with parents/caregivers and/or the patient himself, if applicable [8]. Noteworthy, diagnosis cannot be solely based on rating scales, neuropsychological test or brain imaging. Despite the criticisms that argue a risk of subjectivity, the current consensus supports the validity of the diagnostic criteria applied by well-trained professionals [6]. However, even for a specialist, clinical evaluation is quite time-consuming and requires several visits to be thoroughly performed [9]. Besides, the significant shortage of trained professionals also contributes to a frequent delay in diagnosis or even to overlook some cases. From a developmental perspective, an early diagnosis is very likely to be of value for more effective pharmacological and psychosocial interventions [10]. In this view, there is a need for objective biomarkers as useful adjunctive indicators to alleviate the workload of diagnoses and treatment follow-up.

Numerous studies have tried to assess ADHD through different objective diagnostic tools, most using functional (fMRI) or structural magnetic resonance imaging (MRI) and electroencephalography (EEG), with other modalities (magnetoencephalography (MEG), electrocardiography (EKG), etc) being deployed less frequently, and with an increasing use of artificial intelligence (AI) techniques (for reviews, see [11, 12]). Noticeable efforts in MRI and fMRI were made under the initiatives of the 'ADHD-200 Consortium' [13]. Despite significant advances in understanding abnormalities related to brain maturation and function, neuroimaging findings in ADHD research cannot yet be used to support clinical practice due to a variety of concerns [7, 11]. Likewise, though promising, studies devoted to single-subject prediction of ADHD via deep- and machine-learning (ML) methods have reported quite variable results [12, 14], raising certain methodological concerns [15]. In actual practice, even if useful neuroimaging indicators were available to support clinical decisions, the availability of these diagnostic tools and their associated costs would represent a major barrier to their regular use, further increasing the burden on healthcare resources.

An alternative tool to assess ADHD worth to explore is functional near-infrared spectroscopy (fNIRS), which is characterized by being noninvasive, wearable, cost-effective, and deployable in more friendly/ecological settings (for a review, see [16]). fNIRS has shown its usefulness in monitoring functional hemodynamic changes associated with cortical brain activation (for a review, see [17]). Compared to other neuroimaging modalities, few fNIRS studies have been conducted to discriminate children with ADHD from non-ADHD controls, some of them trying to improve classification by combining different modalities (e.g. EEG + fNIRS) as in [18]. Even fewer studies focused on single unimodal approaches by using 'exclusively' fNIRS data. For example, Monden *et al* reported a classification accuracy of 85% with a sensitivity of 90% by analyzing Receiver Operating Characteristic (ROC curves obtained from right prefrontal oxy-Hb activation data during a go/no-go task [19]. Using prefrontal cortex (PFC) activation measures during an N-back task, Crippa *et al* achieved mean accuracies of 78% with 72% sensitivity and 82% specificity when a support vector machine (SVM) classifier was trained on data from deoxy-Hb [20]. Also employing an N-back task and an SVM, Gu *et al* reached 86% of accuracy with oxy-Hb data measured in the prefrontal and temporal cortex [21]. It is worth noting that no correction for components of non-cerebral origin was applied to the fNIRS signals in the aforementioned studies, which is especially important when scanning the PFC through the forehead [22, 23], since functional extra- and cerebral responses are interrelated processes that overlap in fNIRS recordings and with a greater confounding effect for oxy-Hb [24–26]. Notwithstanding this known drawback of fNIRS, classification algorithms can achieve appreciable performance by learning some type of feature representation from the uncorrected NIRS data, but uncertainty about the nature and origin of the features hampers the interpretability of predictive models. We also note that, in these studies, the features were based on some kind of measurement from the averaged fNIRS data across trials/epochs, a classic approach that, while often providing robust results, fails to uncover finer distinctive patterns embedded in the data.

In a previous study involving non-ADHD young adults, our research team showed that a rhythmic mental arithmetic task successfully induced cyclical hemodynamic fluctuations coupled to the task frequency (33 mHz), and that the oscillatory patterns were consistent across individuals both in superficial and deep fNIRS signals recorded in the frontopolar region [27]. Spectral analysis also showed oscillatory activity at lower frequencies (<33 mHz) seen at rest and during mental task, and with a prominent peak around 5–10 mHz. Resting-state fMRI studies have reported that ADHD patients show significant differences in the low-frequency oscillations (10–80 mHz) band across multiple brain regions [28–30], with separable contribution of specific frequency sub-bands including extra-low frequencies (0–10 mHz) [31]. These differences have been related to abnormalities in the salience, attentional and default-mode networks (DMNs) functioning, but the inconsistences observed across many studies point to a large heterogeneity in spontaneous brain activity in ADHD [32]. Despite this, evidence suggests that some characteristics of ADHD brain activity are sensitive to specific frequency bands.

We hypothesized that a rhythmic mental task might induce/modulate fNIRS-measurable oscillatory activity in particular sub-bands within the 0–80 mHz range, which might be distinctive enough to discriminate ADHD at the individual level. Thus, in the present study, we explored frequency-specific hemodynamic patterns during a cyclical mental arithmetic task administered to a homogeneous sample of 10- to 13-year-old boys with ADHD and non-ADHD controls. A data-driven approach, based on synchronization measures, guided us in selecting the relevant frequencies to extract discriminative features and assess the performance of four commonly used supervised ML methods. Analysis were performed on superficial and regression-corrected deep fNIRS signals recorded from the forehead through a recently introduced multi-distance, multi-channel device [27]. The present work aims to contribute to the objective assessment of ADHD through computer-aided affordable tools deployable in many clinical settings.

2. Method

In this study, we focused solely on fNIRS recordings to find distinctive hemodynamic information by following a data analysis pipeline consisting of: (i) signal preprocessing; (ii) delimitation of the frequency range of interest by spectral power analysis; (iii) identification of relevant sub-bands by time-frequency synchronization analysis; (iv) definition of group-representative oscillatory patterns; (v) computation of features by distance measures; (vi) selection of discriminative features and (vii) classification of individuals using ML linear models.

2.1. Participants

Sixteen clinically referred boys meeting the American Psychiatric Association's DSM-V criteria for the combined ADHD presentation (i.e. inattention + hyperactivity/impulsivity) and 17 age-matched typically developing (TD) control boys were initially recruited. All participants were Caucasian, native Spanish speakers, right-handed, had normal or corrected-to-normal vision and none had other major medical/psychiatric diagnoses. The study was approved by the Ethics Committee of the University Miguel Hernandez according to the Declaration of Helsinki. Written consent was obtained from the parents/guardians of all participants prior to enrollment, none of whom received any financial compensation. Boys diagnosed with ADHD were recruited from a local ADHD support association, while school psychologists referred the control group as TD boys with no history of mental disorder or medication use. Since all ADHD participants were on methylphenidate treatment, they were assessed during the 'drug holiday' period prescribed by the referring clinician, with a minimum four day washout in all cases. One boy from the ADHD group and two from the TD control group were excluded due to NIRS signal quality issues. Therefore, the study ultimately included 15 participants with ADHD (mean age 11.9, SD 1.4, range 10–13 years) and 15 TD controls (mean age 11.6, SD 1.1, range 10–13 years).

2.2. Experimental protocol

In a quiet room, participants sat in a comfortable chair and were asked to relax and keep their eyes on a computer screen 80 cm away. Visual cues and instructions were presented on the screen throughout the experiment. They underwent a slightly modified version of a rhythmic mental arithmetic task described in a previous work [27]. Briefly, the task comprised 10 consecutive 30 s trials, each starting with 15 s of mental calculation followed by a 15 s pause of relaxation. During mental math, participants were asked to iteratively add a small number (5–9) to a three-digit number (100–199) (both numbers randomly chosen), silently and as quickly and accurately as possible. The pause then begins by presenting the question 'Result?' for 5 s, prompting for the voicing of the final result reached, followed by a black screen indicating mental relaxation until a 2 s fixation cross announced the start of the next trial. The task lasted 300 s and was uninterruptedly preceded by 300 s of baseline recording in resting state and followed by another 300 s of recovery in relaxed state (figure 1). To check whether participants engaged with the task well enough, we assessed behavioral performance using two simple scores, total iterations and total exact results achieved throughout the entire task. Note that the 30 s period of the trials corresponds to a frequency of 0.033 Hz, which we will refer to as the task frequency throughout the text. We also note that this design minimizes speech during the task, thus avoiding significant changes in breathing that could affect cerebral hemodynamics [33].

2.3. fNIRS recordings and preprocessing

In this work, we used a newly developed NIRS device (Tehia, Newmanbrain, S.L., Elche, Spain), light and easy to use, which was recently introduced in [27]. In short, it is a multichannel continuous-wave NIRS device that has four emitters and ten detectors arranged in a rectangular patch of 80 × 20 mm, each emitter housing two LEDs at wavelengths 740 nm and 850 nm. Through its duty cycle, the device combines pairs of optodes at different separation distances to

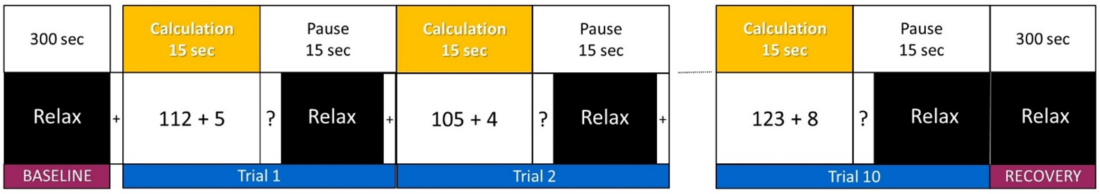


Figure 1. Schematic representation of the experimental procedure.

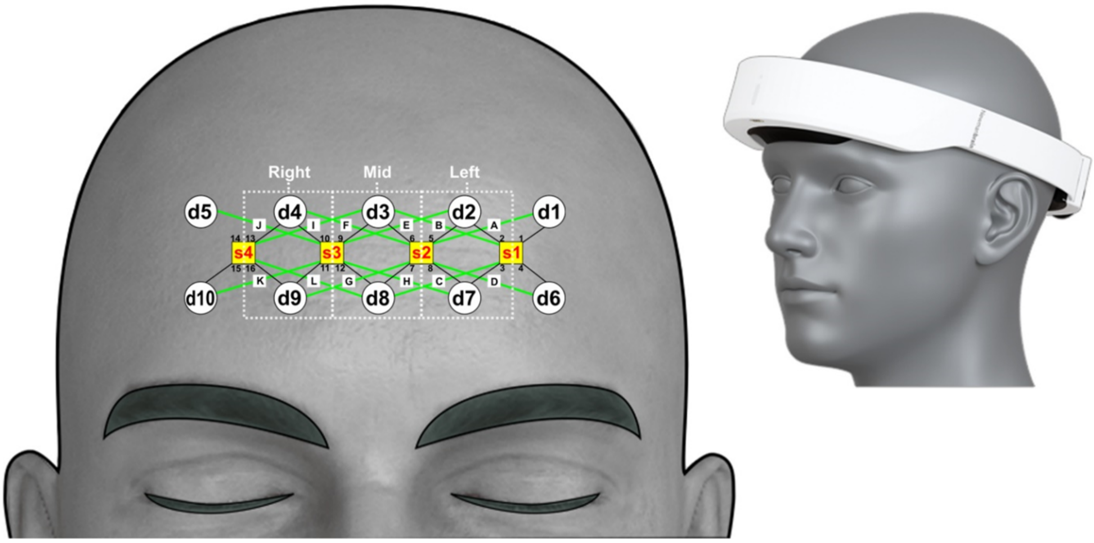


Figure 2. Probe geometry and placement (yellow squares = emitters; white circles = detectors), providing 16 short-channels (black lines with numbers) and 12 long-channels (green lines with letters). White dashed rectangles depict the three regions-of-interests (ROIs) explored in this work.

provide multi-distance recordings through 16 short-channels (14 mm) and 12 long-channels (32 mm). It also corrects for ambient light interference and records motion activity using a 3-axis accelerometer. Data is transferred via Bluetooth at a sample rate of 10 Hz. The NIRS probe was placed on the forehead centered on AFpz according to the international 10-5 system, mainly covering the frontopolar area (Brodmann area 10) of the PFC (figure 2).

To identify recordings suffering from poor signal-to-noise ratio, saturation or unphysiological interferences, we checked the raw optical data to identify those that exhibited extreme values ($<5\%$ or $>95\%$ of the device's dynamic range) or an excessive coefficient of variation ($>7.5\%$) [34, 35]. Furthermore, by visually inspecting sudden changes in signals aligned with sharp shifts in accelerometer data, we identified recordings degraded by motion artifacts. As previously mentioned, three participants were excluded due to signal quality issues.

Relative concentration changes in oxy- (HbO) and deoxy-hemoglobin (HbR) were computed via the modified Beer–Lambert law [36, 37], using functions of the Homer2 NIRS package [38] based in MATLAB (Version R2021b, Mathworks, Natick, MA, USA). A differential pathlength factor of 6 was used for both wavelengths. HbO and HbR data were then digitally low-pass filtered with a zero-phase, 5th-order Butterworth filter, cut-off 0.08 Hz (MATLAB Signal Processing Toolbox); no high-pass filtering was applied. Thus, we remove blood-pressure, respiratory, cardiac and high frequency instrumental components while preserving the band of very-low-frequency oscillations [39]. As a result, for each chromophore we achieved 16 time-series from the short-channels plus 12 from the long-channels that we call shallow- (SSs) and deep-signals (DSs), respectively. SS contains non-cerebral components recorded from superficial layers, whereas DS combines both shallow- and deep-components [24].

Since DS is contaminated by confounding elements unrelated to functional brain activity, such as systemic hemodynamics and inhomogeneous blood flow changes in the superficial tissue layers [22, 26, 40], a cleanup step was necessary to highlight the task-induced cortical response. The NIRS device provides us with multi-distance measurements, where each DS has three candidate SS that can be used as spatially close references to remove contamination [41–43]. As suggested in [44], we applied a 'double SS' approach in which each DS was regressed on the sum of the two SSs recorded closest to the emitter and detector of the corresponding long-channel. We solved linear regression by using the MATLAB function 'robustfit', which uses an algorithm less sensitive to outliers than ordinary least-squares [45]. After regression, the raw residuals represent a clean (or corrected) signal (CS) that likely contains the actual neuronal signal but perhaps mixed with other unpredictable components not fully removed.

Due to variability in head shape and size [46], channel positions are not fully consistent across individuals, and thus interpretation of isolated channels can be misleading. To improve signal reliability by spatial clustering [47, 48], for every single participant we averaged across the SSs and CSs belonging to three regions of interests (ROIs), left, medial and right (figure 2). Therefore, the data for each ROI is now reduced to just two average signals which, for consistency, we will continue to denote as SS and CS. All further processing was done on these signals, which show a comparable signal-to-noise ratio in all three ROIs because they were obtained by averaging the same number of neighboring signals in each region. Finally, to operate on the same scale, all signals were standardized into z-scores.

2.4. Spectral power distribution during task

Cognitive tasks may affect fNIRS signal components at multiple frequencies, reflecting the interlinked contribution of multiple factors arising from systemic, superficial tissue layers and neuronal activity [34, 49, 50]. As a preliminary step, we aimed to identify the most relevant oscillatory components present during the math task, regardless of whether they were spontaneous or task-evoked. To this end, we conducted a power spectral density (PSD) analysis using the Welch's averaged periodogram method [51]. PSD was obtained from the signal segment between 30 s before the task onset and 30 s after the end of the task, which yields a data vector of 30 + 300 + 30 = 360 s in length (3600 samples at sampling rate = 10 Hz). Additional 30 s on both sides were included to account for potential anticipatory or persistence task-related effects. To improve frequency resolution, we extended the data to the next power-of-two (4096 samples) by symmetrical reflection. PSDs were then computed via the MATLAB 'pwelch' function with FFT length = 4096, Hanning window = 2000 samples and overlap = 80% to account for spectral smoothness and reduced noise variance [52]. To allow comparisons, the PSDs were normalized to relative percentage values by calculating the power ratio of each frequency bin to the total power of the entire spectrum [53]. This procedure was applied to the SS and CS data of each participant.

To assess significant PSD differences between TD and ADHD, we performed a two-sample t-tests along frequency bins. The observed t-statistics were corrected for multiple comparisons following the cluster-based nonparametric approach given in [54]. We computed Monte Carlo cluster tests over 2000 permutations of the same t-test by randomly shuffling the data between classes. Then, we estimated the permutation p-value from the proportion of random realizations showing a larger cluster-statistic than the observed one. We set a critical alpha-level = 0.01 to identify frequency bins significantly different. PSDs were finally averaged across the participants of each class to obtain the average normalized PSD of HbO and HbR for each signal type and ROI. The 95% confidence interval (CI) for the mean at each frequency bin was also calculated by bootstrapping over 2000 resamples.

2.5. Time-frequency decomposition

Identification of the frequency components with potential capacity to discriminate between the two classes of participants (i.e. TD and ADHD) was a crucial issue. To this end, we needed a suitable method for locating task-related oscillations on different time scales (i.e. frequency bands), appropriate for non-stationary signal analysis, and capable of providing some measure of similarity to define class membership. Conventional spectral analysis can resolve how the signals' power is distributed along frequencies, but reveals little about how frequency content changes over time or time-varying patterns. Time-frequency analysis (TFA) techniques are a better option for discovering oscillatory patterns at certain frequencies, particularly when these may vary substantially over time and/or multiple components are present in the signal [55, 56]. Furthermore, they allow for separating the magnitude and phase components associated with the signal, which is very useful for capturing transient oscillations alignments [57, 58]. As discussed in [59], the selection of an optimal TFA technique depends on the knowledge of the signal characteristic, which allows a better match with the properties of a particular analysis method. We had no prior information on the most useful oscillatory components for classification purposes, and thus the conventional bandpass filtering + Hilbert transform method was unreliable due to the likely misidentification of the passbands to apply; apart from other inherent limitations [60]. We also avoided signal-adaptive methods that greatly rely on specific knowledge of their parameter settings to find meaningful results, such as empirical or variational mode decomposition [61, 62]. Hence, we decided to use a data-driven approach based on complex continuous wavelet transform (CWT) and time-scale synchronization detection.

CWT is a signal processing method that provides a time-frequency (or time-scale) representation of the characteristics of a signal on the basis of the dilation and translation of a mother wavelet function;

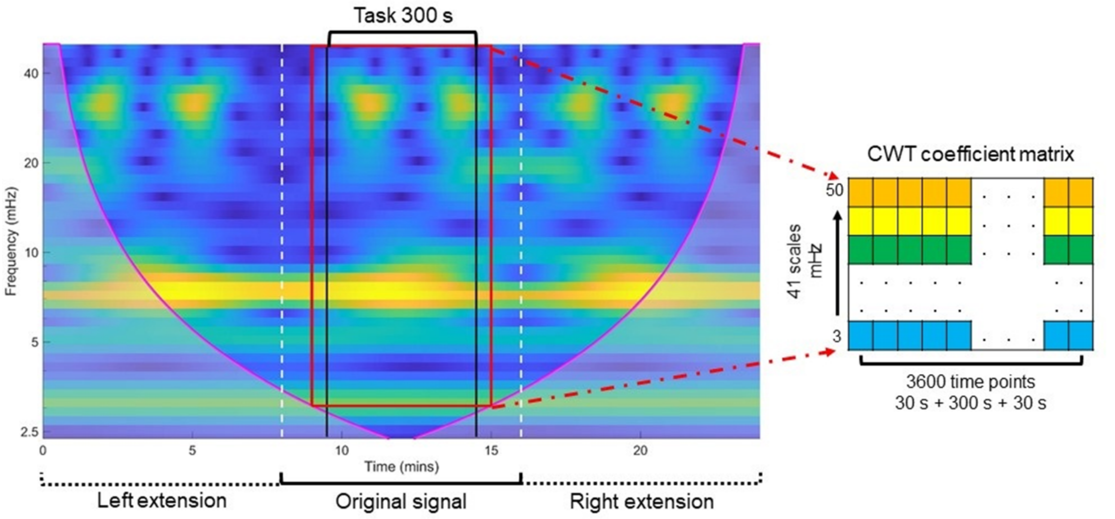


Figure 3. Continuous wavelet transform (CWT) of deep HbO signal from mid-ROI for a representative participant. (left) Scalogram of the symmetrically extended data with the magenta traced cone-of-influence. White dashed lines enclose the original signal and the solid black lines delimit the task-period. The red box depicts the scalogram region from which the coefficients for analysis were extracted. (right) Schematics of the extracted coefficient matrix containing 41 scales and 3600 time points.

theory and mathematical description can be found elsewhere [63, 64]. CWT can be viewed as a bandpass filter with varying bandwidths automatically defined by the wavelet scale [65], which avoids the drawbacks of using custom filters [60]. To compute the CWT we made use of the generalized Morse wavelets, a flexible superfamily of exactly analytic wavelets particularly useful for analyzing signals with time-varying amplitude and frequency, i.e. modulated signals [66, 67]. Since Morse wavelets can be tuned to encompass many other analytic wavelets commonly used, they provide a unified framework as reference point. Thus, for example, setting the symmetry parameter $\gamma = 3$ defines a family members (‘Airy’ wavelets) that can successfully replace the popular Morlet wavelet as the default analytic wavelet for general-purpose use [68]. Besides other applications, Morse wavelets have been proposed in a variety of biomedical studies, such as human locomotion [69], electrocardiography and electromyography [70, 71], neural oscillations coupling [60, 72], electroencephalographic data classification [73, 74] and fNIRS artifactual denoising [75].

To keep the procedure as simple and reproducible as possible, we computed the CWT using the default Morse parameters recommended in the MATLAB Wavelet Toolbox (symmetry = 3, time-bandwidth product = 60), which yield a perfectly symmetric wavelet. To obtain the most accurate time-frequency representation of our time-series data, we accounted for edge effects produced when the wavelets extend (‘see’) outside the signal boundaries. Since such effects depend on the wavelet scale, the so-called ‘cone of influence’ (COI) must be considered to avoid possible inaccurate coefficient values. Of note, in the case of Morse wavelets the equivalent concept is ‘wavelet footprint’ [76]. In order to reduce edge effects, we extended the signals through symmetric reflection of the full signal length prior CWT.

As the preliminary PSD analysis (section 2.4) showed that most of the spectral power was within the 0–50 mHz band (figure 6), we focused the CWT on that frequency range. CWT was applied to the symmetrically extended signals with the scale discretization parameter voices-per-octave = 10, which after calculation of the minimum and maximum allowable bandpass results in 45 scales with approximate frequencies ranging from 2.4 to 50 mHz. Despite signal extension, we observed that frequencies below 3 mHz were slightly outside the COI boundaries (as estimated by the MATLAB ‘cwt’ function) and more prone to edge effects, so we exclude them from subsequent analysis (figure 3). Thus, the number of usable scales was limited to 41 (3–50 mHz), which in turn excludes any extremely slow trends that were not removed during signal preprocessing.

After CWT, we kept the full coefficient matrix for later computation of the inverse CWT (i.e. including the data corresponding to the extended segments) whereas for the next step we would use only a shorter portion of the matrix.

2.6. Time-scale synchronization detection

Under the hypothesis that the math task may induce differentiated fNIRS fluctuations for the TD and ADHD groups, we set out to identify the frequency sub-bands that showed higher group synchronizations during the task. Since group-wise synchronization may appear as transient peaks rather than constantly, we performed a time-point-by-time-point

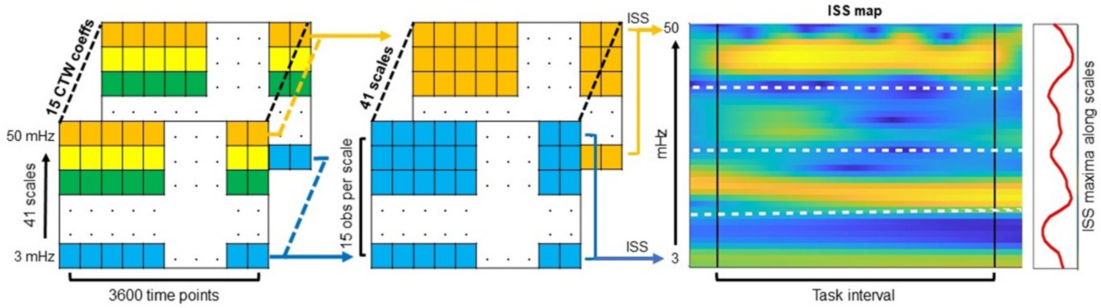


Figure 4. Schematics of the procedure to compute a representative inter-subject synchronization (ISS) map. Starting with the coefficient matrices of each individual within a group (left), complex-valued data are put together for each frequency bin (middle) and the ISS is computed moment-to-moment to obtain a representation of the within-group synchronization strength at each frequency over time (right colormap). The rightmost red plot depicts the maxima and minima reached by the ISS along frequencies, which are used to delimit the sub-bands to analyze (horizontal white dashed lines).

analysis, which allows capturing common oscillatory patterns that evolve dynamically over time. It is worth noting that this strategy is inspired by the underlying logic of so-called inter-subject correlation analysis, a data-driven approach devoted to assessing consistent neural responses to stimuli across individuals [77, 78].

We measured instantaneous inter-subject synchronization (ISS) using the magnitude and phase information provided by the complex-valued CWT coefficients. In fMRI studies, measures as inter-subject phase synchronization [58] and pairwise phase consistency [79, 80] have been validated for the assessment of voxel-wise instantaneous phase synchronization across subjects. The former is similar to the 'circular mean resultant length' in circular statistics literature [81], also known in EEG event-related studies as 'inter-trial phase coherence' [82], 'inter-trial phase clustering' [83], or 'phase-locking factor' [84]. However, these measures rely only on the uniformity of phase angles, ignoring the magnitude. Thus, when applied to fNIRS it means that low-amplitude signals affect the measurement the same as those with significant amplitude. Therefore, this approach may not be entirely appropriated for fNIRS data where amplitude changes are related to the magnitude of the hemodynamic response. Since as amplitude increases, the signal-to-noise ratio improves, it is reasonable to argue that observations with higher amplitudes can contribute to a more realistic estimate of phase synchronization [85]. Under this assumption, we decided to use a closely related measure called 'inter-trial linear coherence' that combines magnitude and phase in the normalization step [82]. Since the measurement here was across subject observations and not across trials, we call this ISS, omitting 'linear' for simplicity and similarly formulated as:

$$\mathrm{ISS}\left(f,t\right) = \frac{\sum_{k=1}^{n} F_k\left(f,t\right)}{\sqrt{n\sum_{k=1}^{n}\left|F_k\left(f,t\right)\right|^2}},$$

where $F_k\left(f,t\right)$ is the spectral estimate of observation k at frequency f and time t, and the modulus $||$ represents the complex norm. ISS also takes values between 0 (absence of sync) and 1 (perfect sync).

ISS was computed moment-to-moment for each scale from the CWTs coefficient matrices of each group (i.e. 15 participant observations per scale). As done in section 2.4, we limited the analysis only to the time-interval between -30 s and $+30$ s around the task. This procedure was applied independently to the SS and CS data of each group of boys for each chromophore and ROI, obtaining an ISS representation in the time-frequency plane that can also be visualized as a color map (figure 4). Next, we choose the maximum ISS observed along each scale, which represents the highest group synchronization achieved at each specific frequency (figure 4). Please note that in this work we only take into account the ISS maxima, regardless of the time point at which they are reached. We also did not consider calculating any significance threshold at this step, as we relied on the inherent discriminative power of the subsequent feature selection and classification procedures.

Based on the observed maxima, on each ISS map we identify several peaks within specific frequency sub-bands that contain oscillatory components showing some synchronization at the group level. Since we expected such components to provide information to differentiate between groups, it was necessary to locate the sub-bands exhibiting more discriminative power. To simplify the procedure, for each case we averaged the ISS data across the three ROIs to obtain the mean ISS maxima per frequency. Thus, we simplified the analysis to only one common ISS pattern by chromophore and signal type for each of the two groups, i.e. 2 groups (TD, ADHD) × 2 chromophores (HbO, HbR) × 2 signal types (SS, CS) = 8 ISS patterns. Finally, to reduce noise, the ISS patterns were smoothed by moving average using a sliding window of length half the voices-per-octave (i.e. $10/2 = 5$). These patterns were examined in the following steps to identify the most relevant frequency components.

2.7. Determination of frequency sub-bands and feature extraction

Like the other aforementioned synchronization measures, ISS is a compound measure that does not exist on its own at a single-subject level but represents a summary statistic of group synchronization. Therefore, to disentangle the contribution to ISS of each individual is not a straightforward issue [86]. However, ISS peaks suggest that some frequency components show similar time courses across individuals, at least within certain time intervals. In other words, there are sequential patterns common to the group that can provide distinctive information to define class membership. This concept falls into the interdisciplinary and much-studied field of time-series classification, which encompasses a variety of techniques for identifying those properties (features) that have sufficient discriminating power to distinguish between different classes of time series (for a review, see [87]). In the context of the present work, a well-suited technique could be the one based on the *shapelet* framework, which addresses the classification problem by discovering primitive time-series sequences (*shapelets*) that are used to quantify the (dis)similarity between classes of time series [88, 89]. *Shapelets* provide directly interpretable information about patterns (shapes) that are important for understanding how data classes differ, a desirable property for clinical decision support systems [90].

Here we applied the basics of the *shapelets* technique, but instead of looking for phase-independent subsequences similar in shape (i.e. subsequences may be located anywhere in the series) we performed the analysis within a fixed time-interval, all subsequences having the same length. We did not apply any subsequence translation over time, which implies that time-series similarity also depends on the phase (i.e. on a consistent time-alignment). Therefore, instead of local, we captured global patterns present over a whole time interval. Under this approach, the term '*shapelet*' may not be appropriate; however, since it relies on comparable principles and is easy to conceptualize, we will keep it here but in terms of a pattern that is maximally representative of a class within a specific timeframe.

At this point, we need to extract the time-series to be used for identifying representative *shapelets*. The average ISS patterns suggest us the frequencies that are likely to contain synchronized oscillations. By computing the inverse CWT within the specific sub-band defined by the bounds of an ISS peak, we can reconstruct such band-limited components in the time-domain. To reduce edge-effects, the inverse CWT was computed from the extended coefficient matrix that we reserved in a previous step (see section 2.5). Then, the resulting time-series were truncated to the interval between −30 s and +30 s around the task. After applying this procedure to the CWT of all the individuals belonging to a group, a set of time-series ($n = 15$) is available to find a reference *shapelet* for that group in a particular sub-band.

Since all the time-series have the same length and are within the same timeframe, a suitable reference *shapelet* can be obtained simply by averaging. If the time-series share a common pattern, their average should represent the group well enough. To quantify similarity with the reference *shapelet*, among other possibilities, a simple measure as Euclidean distance can be computed:

$$D(S,T) = \sqrt{\sum_{i=1}^{n} (S_i - T_i)^2}$$

where S denotes the reference *shapelet* and T a time-series, both of length n. Note that S and T should be standardized to have mean 0 and standard deviation 1, which ensures to operate on the same scale. Standardization also allows us to relate the Euclidean distance to the correlation coefficient ($r = 1 - D^2/2n$).

In order to assess the capability of the *shapelet* to discriminate between groups, we can contrast the distances obtained from one group with those of the other group. For example, a *shapelet* that is representative of the TD group should have smaller distances to members of this group than to members of the ADHD group, and vice versa. Among other quality measures, the F-statistic for analysis of variance can be used to assess the discriminative power of a *shapelet* [89]. This statistic indicates the ratio of the between-group variability to the within-group variability as:

$$F = \frac{\sum_{i=1}^{C} \frac{(\bar{D}_i - \bar{D})^2}{C-1}}{\sum_{i=1}^{C} \sum_{d_j \in D_i} \frac{\left(d_j - \bar{D}_i\right)^2}{n-C}}$$

where C is the number of classes (or groups; in our case = 2), $\bar{D}$ is the overall mean of all distances, $\bar{D}_i$ is the mean of the distances within class i, and n is the number of time-series. The better the *shapelet* the greater the F value, because the difference between-groups increases while it decreases within a group. The corresponding p-value can be calculated from the F-distribution.

Based on the average ISS patterns, we identified four candidate sub-bands in each of them that were labeled *A*, *B*, *C* and *D* in decreasing order of frequency. Each sub-band contains a peak (local-maximum) flanked by two troughs (local-minima) that delimit the frequency boundaries. For each ISS pattern, each belonging to a target group (TD or ADHD), we performed the following procedure for each sub-band (figure 5): (i) compute the inverse CWT within the sub-band from the extended coefficient matrices of both groups to obtain the corresponding time-series ($n = 15 + 15 = 30$). (ii) Truncate time-series to the time-interval of interest. (iii) Generate the reference *shapelet* by averaging only the

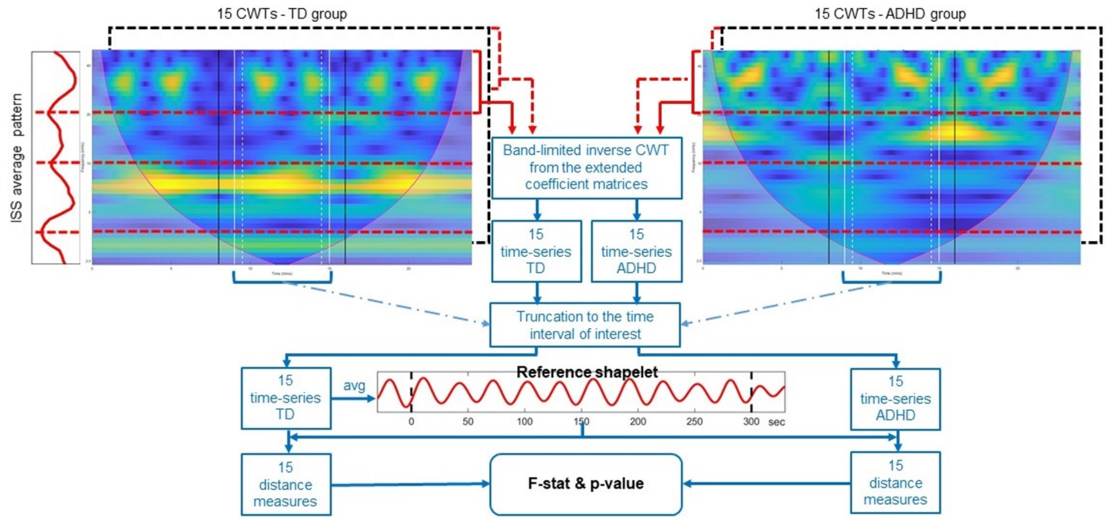


Figure 5. Schematic example of the procedure to extract reference shapelets for the TD group and contrast them with the ADHD group. Per-participant time-series within a specific frequency sub-band are obtained from their band-limited inverse-CWT and averaged only for the TD group to get the shapelet. Euclidean distances to the shapelet are computed for the time-series of both groups and then within- and between-group variances are contrasted to obtain the F-stat value.

time-series of the target group ($n = 15$), a variability measure such as the standard error of the mean can also be computed. (iv) Calculate distances to the *shapelet* for the time-series of both groups. (v) Calculate F and p-value from the $15 + 15 = 30$ distances. Because each ISS pattern and its sub-bands are common to all three ROIs, this procedure was applied separately to each ROI data but using the same sub-bands. Thus, for each ISS pattern we obtain a matrix of 30×12 distances, where each row contains the distance measures of an individual and columns correspond to the 4 sub-bands × 3 ROIs. Since there are eight ISS patterns, the final matrix was of size 30×96, 48 columns for SS and 48 for CS, while 15 rows correspond to the TD group and 15 to ADHD. Please note that each column has an associated F-stat and p-value, indicating how well a particular *shapelet* differentiates the groups.

In summary, we employed the *shapelet* approach to transform data observations at different time-scales into a simple feature space of Euclidean distances, which are the only feature type used in the present work.

2.8. Classification algorithms and feature selection

To assess the feasibility of the proposed procedure to differentiate between TD and ADHD, we tested it with four well-suited machine learning algorithms for supervised binary classification, namely linear SVM [91], logistic regression (LR) [92], linear discriminant analysis (LDA) [93] and Gaussian naïve Bayes (NB) [94]. We chose these algorithms because they are well known, inherently interpretable, computationally efficient, and can work with relatively small sample sizes. Under a variety of flavors (different kernel, regularization, etc), SVM is very frequently present in neuroimaging-based studies of brain disorders, with LDA and LR being the other most popular choices [14]. Noteworthy, a similar usage scenario occurs in the ADHD research field [12]. Although less commonly used, we included NB because its ease of application and good performance in a variety of applications despite the assumption of feature independence [95].

Although based on different models, discriminative (LR & SVM) vs generative (LDA & NB), all four are within the linear classifier category, i.e. to make predictions, the classifiers try to learn the line that best separates the points of the two classes [96, 97], which depends on a linear combination of the explanatory variables. Thus, it is possible to know to what extent each feature influences the prediction, which greatly improves its interpretability. In this sense, we avoided more complex classifiers such as those using artificial neural networks or non-linear kernels, wherein the relationship between features and prediction is less transparent. We used MATLAB's implementations of the classifiers (Machine Learning Toolbox) with default settings for simplicity and reproducibility. To minimize the risk of the algorithms overfitting the available data and losing generalizability, hyperparameter optimization was not applied [98]. As usually recommended, the features were standardized using the corresponding column mean and standard deviation.

In addition to the putative functional response, fNIRS signals also contain components originating from common systemic forces and unpredictable local activity. Therefore, it is very likely that our feature matrix also contains redundant and/or irrelevant data that can degrade classifier performance by cause of overfitting and noise issues. Model regularization can be applied to some algorithms to account for statistical overfitting, however that raises the problem of choosing a suitable technique (e.g. *lasso*) and finding appropriate regularization parameters. To avoid increasing the complexity of the models, we addressed the problem by reducing the feature space. Feature selection is a commonly used tool to obtain a smaller subset of the most relevant features, reducing complexity while improving classification accuracy and generalization capacity [99]. A reasonable hybrid approach is to first apply a filter method, before modeling, to select some features based only on their intrinsic properties. Then more sophisticated methods such as *wrappers* may be employed to find the best subset of features, using the classifier itself as evaluator [100]. Among others, a benefit of such selection is an easier explanation of the prediction because the models are simpler [101].

Despite the ample offer of filter methods [102], and after trying some of the popular ones (*relief, minimum redundancy-maximum relevance* and *chi-square*; results not shown), we settled on a fairly straight option based on the *F*-statistic. We simply selected the features that showed p-values $< .01$, assuming that their generating shapelets were very unlikely to separate the groups by chance. In this way, we significantly reduced the features from 48 to 5 for SS and from 48 to 10 for CS.

The classification methods were first applied to filter-selected features separately for SS and CS, and then for all of them together (SS + CS). To assess the predictive performance, we applied two cross-validation (CV) techniques for comparison purposes, namely leave-one-out (LOO) and stratified 5-fold. In the first, data was partitioned into 30 folds where each observation was used once as a test set and the remaining ones formed the training set. In the latter, five partitions were randomly chosen, each with 24 observations as the training set and 6 as the test set; folds were repartitioned over 20 *Monte-Carlo* repetitions ($5 \times 20 = 100$ models) to reduce CV variance [103], while stratification ensured that sets had the same proportion of classes (50% in our case). We used 5- instead of 10-fold because with the latter the test set size = 3 would be too close to that of LOO = 1. Since we are dealing with only two classes and our datasets are well-balanced (i.e. equal proportion of both classes), accuracy, specificity and sensitivity can be used as metrics to assess performance [104, 105], as obtained from the corresponding confusion matrices and then averaged across folds. At this point, we focused on accuracy (a commonly used metric in practice) to test the statistically significant classification performance. Thus, we computed the theoretical above-chance accuracy threshold based on the binomial cumulative distribution at $p < 10^{-3}$ for 2-classes (probability = 0.5) and a sample size = 30 [106].

Afterwards, we applied a *wrapper* method to fine-tune the feature selection. We used a custom-made MATLAB *wrapper* function that implements a sequential forward floating selection (SFFS) algorithm [99, 107]. SFFS starts with an empty set and sequentially adds one feature at a time to create candidate subsets that are evaluated by CV. After that, the best feature is added to the set. When the size of the selected set is >2, a backward step tries to optimize it by removing one or more features. This procedure is repeated until there is no performance improvement. The two aforementioned CV techniques were also applied independently to each classifier. Noteworthy, the input order of the finally selected features allows us to know their relative importance. It is also worth mentioning that our *wrapper* can rank features by multiple metrics at the same time, and that in this work we used two criteria to select/remove features, specificity and then accuracy. Thus, if two (or more) features equally improve the specificity, the one with the best accuracy is selected.

Once the best subset of features was selected for each classifier, we estimated the statistical significance of the observed performance through a non-parametric label permutation procedure that does not assume any particular statistical property of the data [106, 108, 109]. We generated 5000 resamples, each of which randomly permuted the labels of the two classes; realizing the null hypothesis that features do not define class membership. For each resampled data, the classification performance was evaluated using the same CV scheme as for the real data. The observed performance metrics were ranked against the corresponding null-distribution to estimate a p-value. In addition, a 95% bias corrected percentile interval was estimated as CI for each metric by bootstrapping (with replacement) over 2000 resamples, with each realization keeping the same proportion of classes (50%) and at least three distinct observations in each class [110].

Finally, we checked whether the *wrapper* performance might have been biased due to the use of a pre-filtered feature set that included all data in the selection process, i.e. the 'peeking' effect [15, 111]. To this end, we repeated the *wrapper* procedure but using the full feature set (i.e. 96 features), then comparing the performance outcomes.

3. Results

3.1. Behavioral performance

On average, TD participants achieved slightly higher scores for iterations ($M_{\mathrm{TD}} = 36.20$; $\mathrm{SD}_{\mathrm{TD}} = 16.48$; $M_{\mathrm{ADHD}} = 29.66$; $\mathrm{SD}_{\mathrm{ADHD}} = 14.01$) and precision of results ($M_{\mathrm{TD}} = 7.26$; $\mathrm{SD}_{\mathrm{TD}} = 1.79$; $M_{\mathrm{ADHD}} = 7.00$; $\mathrm{SD}_{\mathrm{ADHD}} = 2.32$) than the ADHD group. However, the unpaired t-test did not show significant differences in iterations ($t(28) = 1.169$; $p = .252$) or in precision ($t(28) = 0.351$; $p = .728$).

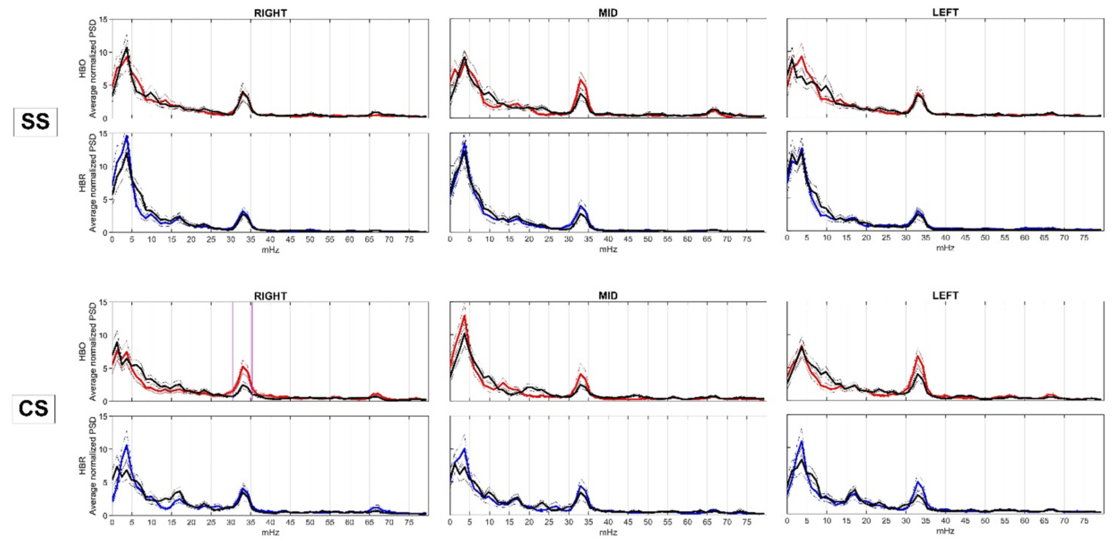


Figure 6. Average normalized power spectral density (PSD) for each ROI (right, medial and left), chromophore and signal type in the frequency range 0–80 mHz. Thick colored curves refer to TD group (HbO in red and HbR in blue) while thick black ones correspond to ADHD group. Thin lines depict the 95% CI of the mean. Magenta rectangles delimit the frequency ranges that show significant differences between groups. Panel SS shows the PSDs for shallow-signals and CS for clean-signals.

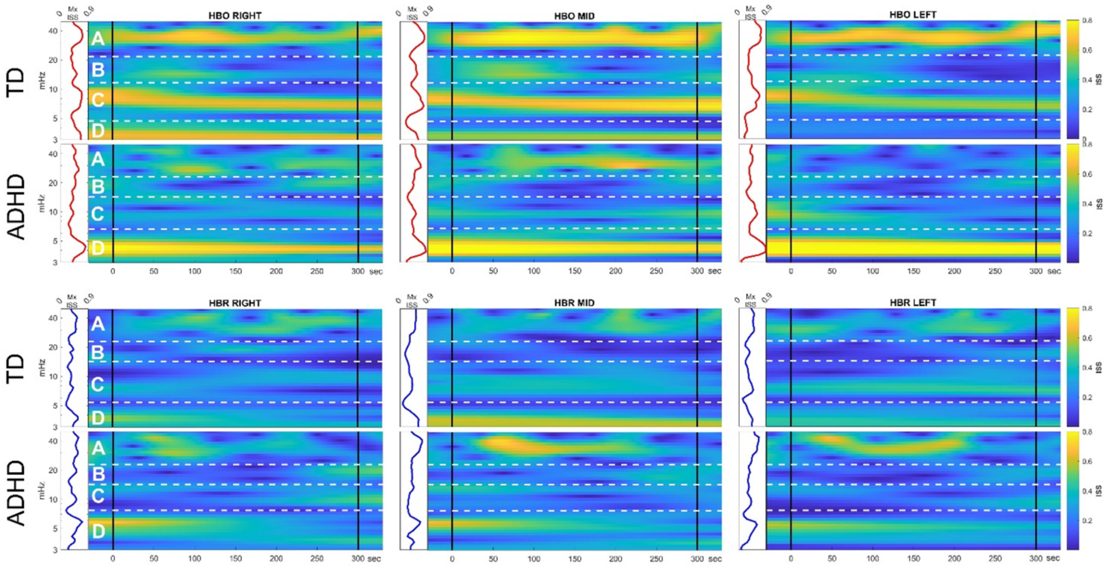


Figure 7. Inter-subject synchronization (ISS) color maps for shallow-signals (SS) within each ROI. Upper rows correspond to HbO for TD and ADHD groups, while lower rows refer to HbR. The small plots to the left of each map depict the ISS maxima across frequencies for each case. Vertical black lines delimit the task-interval. Horizontal white dashed lines delimit the common sub-bands obtained by averaging ISS maxima across ROIs. Labels (A)–(D) identify each of these sub-bands.

3.2. Spectral power analysis

Figure 6 shows the average PSD distribution of fNIRS signals for the TD and ADHD group within each ROI. A common pattern can be observed, with a peak around 33 mHz corresponding to the task frequency and a most prominent peak around 4 mHz. Smaller secondary peaks are also present around 17, 24 and 67 mHz, for example. The cluster-based permutation test only showed significant differences between groups at 33 mHz for CS-HbO in the right-ROI. In all cases, most of the spectral power can be allocated roughly within the 0–50 mHz frequency range.

3.3. Time-frequency ISS maps

Figure 7 shows the ISS representation in the time-frequency plane obtained from the complex CWTs of the SS data for both groups; within the range of 3–50 mHz.

Similarly, figure 8 depicts the ISS maps corresponding to the CS data. At first glance, well-defined synchronization zones can be seen within certain frequency sub-bands, some similar in both groups and others clearly differentiated. Here we highlight some of them as examples. Regarding SS, strong synchronization can be seen in all ROIs around 33 mHz for

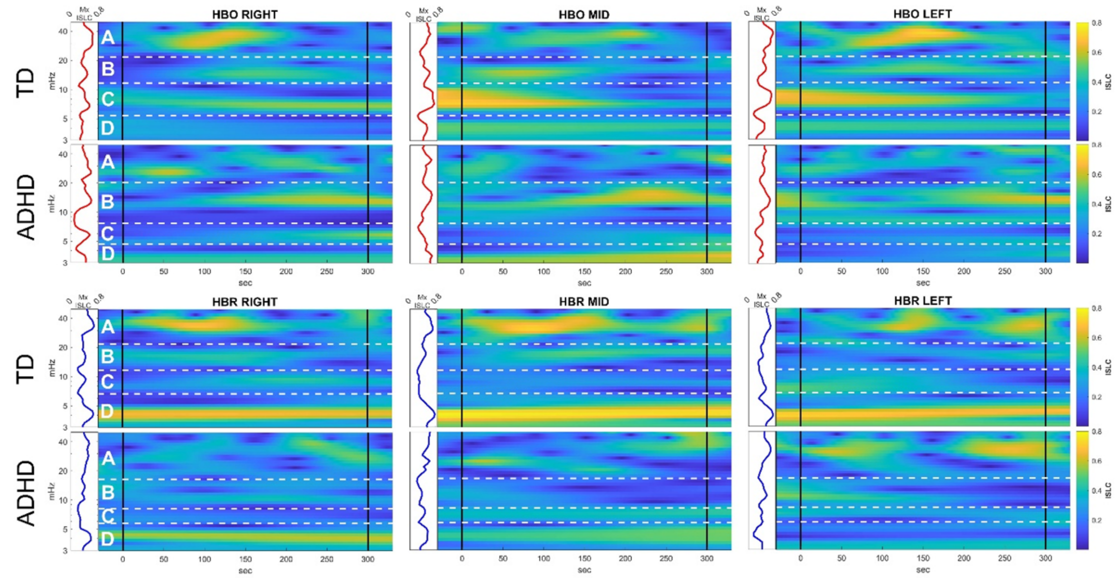


Figure 8. Inter-subject synchronization (ISS) color maps for clean-signals (CS). The description of the drawn elements is the same as in figure 7.

Table 1. Frequency bounds of each average sub-band A–D for shallow-signals (SS) and clean-signals (CS), and for each chromophore and group.

Signal type	Chromophore	Group	Frequency sub-band (mHz)			
			A	*B*	*C*	*D*
SS	HbO	TD	50–21.8	21.8–11.7	11.7–4.7	4.7–3.0
		ADHD	50–23.3	23.3–14.3	14.3–6.7	6.7–3.0
	HbR	TD	50–23.0	23.0–14.3	14.3–5.4	5.4–3.0
		ADHD	50–23.0	23.0–14.3	14.3–7.7	7.7–3.0
CS	HbO	TD	50–21.8	21.8–11.7	11.7–5.4	5.4–3.0
		ADHD	50–20.3	20.3–7.7	7.7–4.7	4.7–3.0
	HbR	TD	50–21.8	21.8–11.7	11.7–6.7	6.7–3.0
		ADHD	50–16.5	16.5–8.2	8.2–5.8	5.8–3.0

HbO of TD group, and in right- and mid-ROI at 7 mHz and below 4 mHz. Noteworthy, the ADHD group presents even stronger ISS around 4 mHz in all ROIs, but much less evident at 7 or 33 mHz. Regarding CS, albeit to a lesser extent, TD group also synchronizes at 33 mHz while TD group does so in a more diffuse and weak way. Furthermore, TD group seems to be more synchronized during the first part of the task at 7 mHz (mid- and left-ROI), whereas the ADHD group is synchronized in the last part around 17 mHz. Yet another remarkable sync is observed for the HbR of TD group at 4 mHz. Overall, ISS analysis reveals a plurality of sub-bands that can carry information about similarities and differences between groups.

3.4. Frequency sub-bands and reference shapelets

Table 1 shows the common synchronization sub-bands estimated from the average ISS maxima across ROIs, labeled *A*, *B*, *C* and *D* by decreasing frequency, with *A* corresponding to the task frequency. Figure 9 illustrates how these sub-bands were delineated by locating the ISS minima surrounding each peak. Higher peaks can be seen in sub-bands *A* and *C* for HbO of TD group in both SS and CS, while ADHD group shows notable peaks in *D* for SS and *B* for CS. Regarding HbR, it shows clear peaks in *A* and *D* in all cases. Note that within the same assigned sub-band, in some cases the peaks are clearly shifted in frequency depending on the group (e.g. *C* sub-band for CS-HbO). It is evident again that the ISS maxima also reveal differences at certain frequencies. These average sub-bands are also depicted in the ISS maps of figures 7 and 8 by white dashed lines.

Figure 10 shows the reference *shapelets* obtained in each sub-band for SS data and figure 11 those corresponding to CS data. A rich variety of patterns can be seen, some similar across groups and others clearly different. Thus, for example, TD group exhibits rhythmic fluctuations in the *A* sub-band of SS-HbO, which are very consistent across participants as reflected by the high ISS (dashed traces); in contrast, ADHD group shows greater inter-subject variability. Another example is visible in CS-HbR-D, were TD

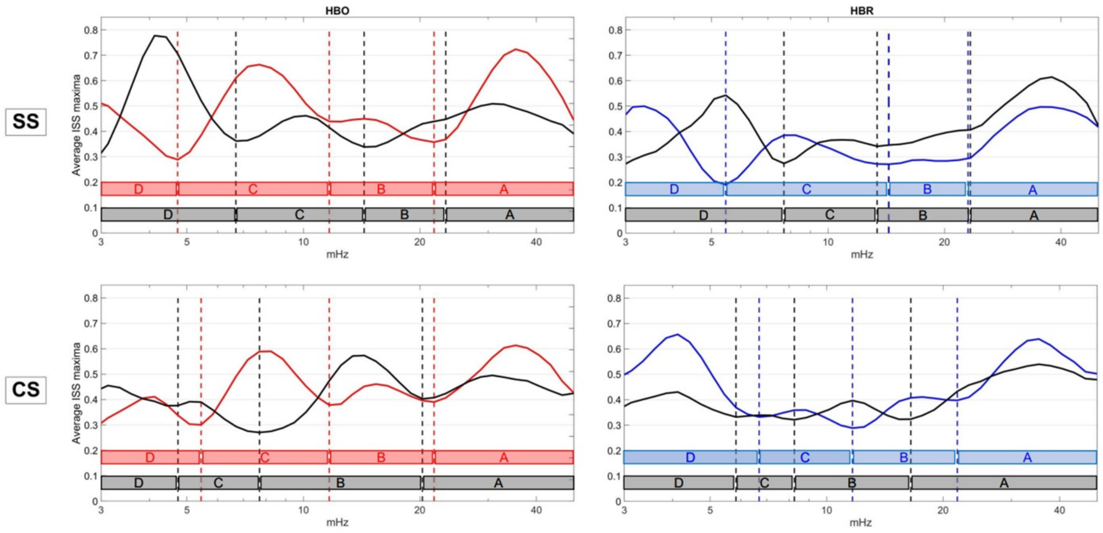


Figure 9. Average inter-subject synchronization (ISS) patterns across frequencies. The top row corresponds to shallow-signals (SS) and the bottom row to clean-signals (CS). Red and blue traces relate respectively to HbO and HbR of the TD group, while ADHD data is always plotted in black. Horizontal labeled rectangles identify each sub-band, whose boundaries are defined by the minima marked by vertical dashed lines.

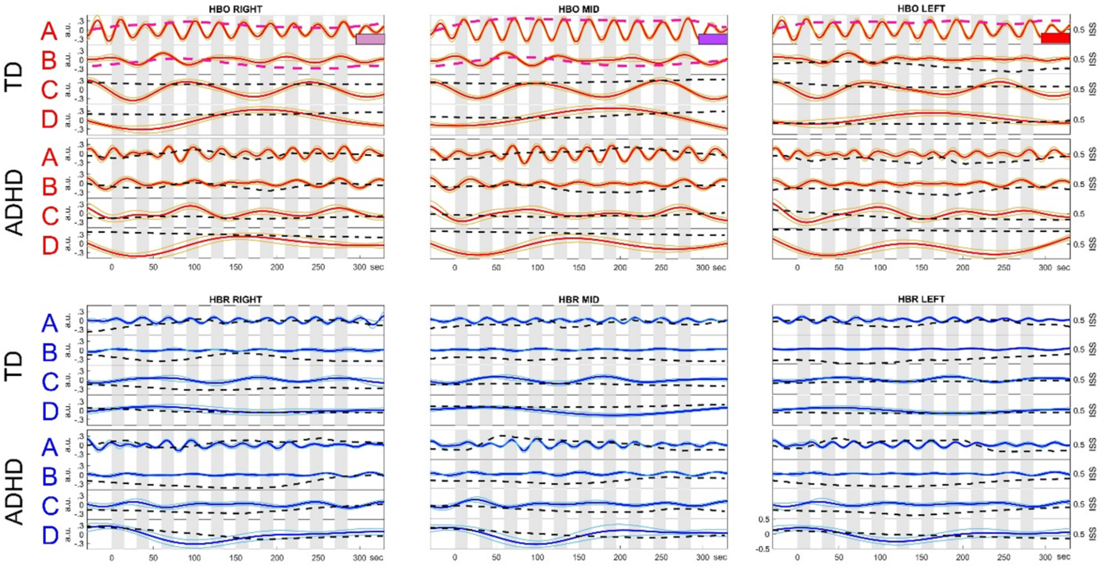


Figure 10. Shallow-signals (SS) reference *shapelets* for each sub-band (*A–D*) within each ROI. Upper rows show the HbO patterns (red traces) for TD and ADHD groups, while lower rows refer to HbR (blue traces); thin traces depict the *shapelet's* SEM. Dashed lines show the inter-subject synchronization (ISS) time-course in each sub-band (scale on the right axis); in magenta color those corresponding to *shapelets* with an *F* *p*-value < .01. The colored boxes identify the *shapelets* chosen by the *wrappers* (see figure 12). Gray shaded rectangles indicate each of the 15 s of mental math during task.

group shows a consistent pattern of increasing then decreasing, whereas ADHD group does not. It can also be seen that the ADHD group synchronizes CS-HbO-B towards the end of the task, while TD does so visibly earlier. Once again, certain *shapelets* seem to represent well the average response of their group, while they do not fit the other one.

3.5. Classification performance

Table 2 shows the performance achieved by classifiers trained with the features selected by filtering, i.e. those with an *F* *p*-value < .01. Five and ten *shapelets*, respectively for SS and CS, generated features that met the filter criteria, each identified in figures 10 and 11 by the magenta color of their ISS traces. Overall, performance improves for all classifiers when SS and CS features are combined. All of them reached accuracy values ⩾76.7% with both CV schemes, which is the above-chance threshold at $p < 10^{-3}$ according to the theoretical binomial cumulative distribution for 2-classes and a sample size = 30. Regarding to specificity, LDA showed the highest values (88.7% and 93.3%, respectively for five-fold and LOO). When SS and CS features were used

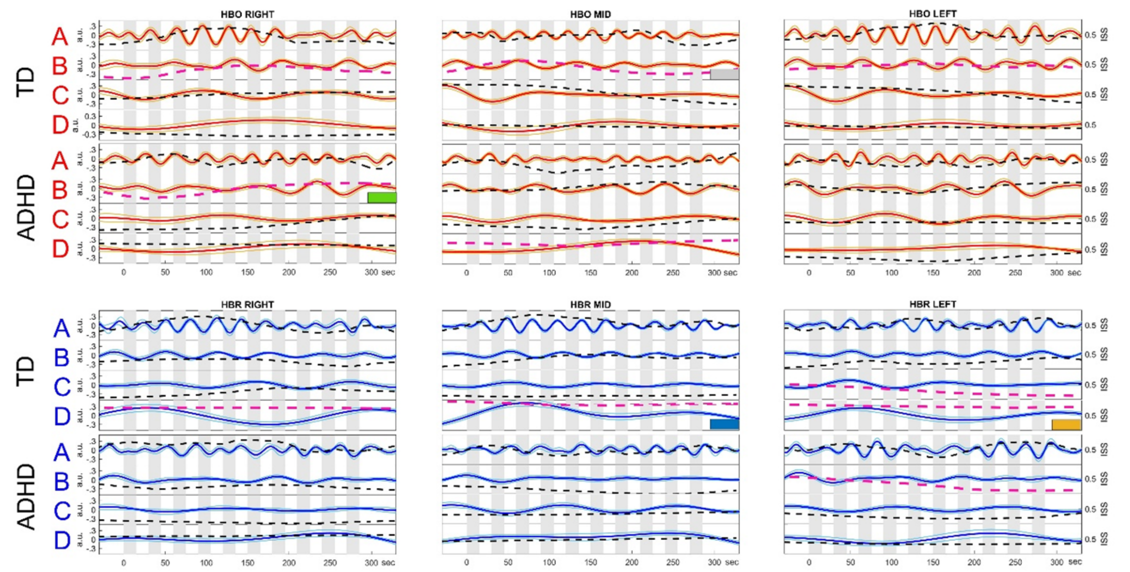


Figure 11. Clean-signals (CS) reference *shapelets*. The description of the drawn elements is the same as in figure 10.

Table 2. Performance scores achieved by each classification model trained with the subset of filter-selected features for shallow-signals (SS), clean-signals (CS) and SS + CS.

		SVM		LR		LDA		NB	
Signal type	P. metric	5-FO	LOO	5-FO	LOO	5-FO	LOO	5-FO	LOO
SS 5 features	Accuracy	73.3	70	80	80	74.5	73.3	80.5	80
	Sensitivity	78.7	73.3	85.7	86.7	80.7	80	88.7	86.7
	Specificity	78.7	73.3	85.7	86.7	80.7	80	88.7	86.7
CS 10 features	Accuracy	81.2	80	83.5	80	78.7	80	85.5	86.7
	Sensitivity	82.3	80	82	80	81.7	80	84.7	86.7
	Specificity	82.3	80	82	80	81.7	80	84.7	86.7
CS & SS 15 features	Accuracy	88.7	86.7	88.2	90	85.7	83.3	85.7	86.7
	Sensitivity	91.3	93.3	91.7	93.3	82.7	73.3	84	86.7
	Specificity	86	80	84.7	86.7	88.7	93.3	87.3	86.7

Note: LOO = leave-one-out; 5-FO = 5 folds; SVM = support vector machine; LR = logistic regression; LDA = linear discriminant analysis; NB = naïve Bayes.

independently, performance was worse, although all classifiers still showed above-threshold accuracies for CS, while only LR and NB did so for SS. These preliminary findings suggest that SS and CS together contain valuable information to discriminate groups independently of any classification model, CS being probably the more relevant.

Figure 12 demonstrates how performance was greatly improved by using the *wrapper* for feature selection. Note that, in all cases, the classification models were very parsimonious and no more than three features were used. When compared separately, *wrapper*-selected CS features perform better than SS overall. Regarding the input order, for SS the first-in feature always belongs to TD group, A sub-band, HbO, left- or right-ROI depending on the classifier ('TD-A-HbO-L' magenta box or 'TD-A-HbO-R' red box in figure 12). It should be noted, regarding CS, that all the classifiers agree on the first two features 'AD-B-HbO-R' and 'TD-D-HbR-L' (green and yellow boxes, respectively, in figure 12). Once again, the best results were obtained with SS + CS. In fact, LR and LDA scored over 99% on all three metrics for both CV schemes, which is really high performance. NB scored lower, from 91.3% to 93.3% overall while SVM was the weakest classifier when all metrics are considered.

When evaluating SS + CS, the *wrapper* selected the same features for LR and LDA in both CV cases. Noteworthy, the first-in feature was 'AD-B-HbO-R', which was also the first for CS. The second one was 'TD-A-HbO-L', the first for SS. Finally, 'TD-D-HbR-M' from CS completed the set. Looking at the *wrapper* history, we observed that 'AD-B-HbO-R' alone achieves about 80% of the accuracy, sensitivity and specificity, which is not surprising since it has the highest F (19.6, p-value $< .0001$). The addition of 'TD-A-HbO-L' improves the scores up to 90% and with 'TD-D-HbR-M' they approach 100%. Therefore, the most powerful feature comes from the CS-HbO data of ADHD group, specifically from

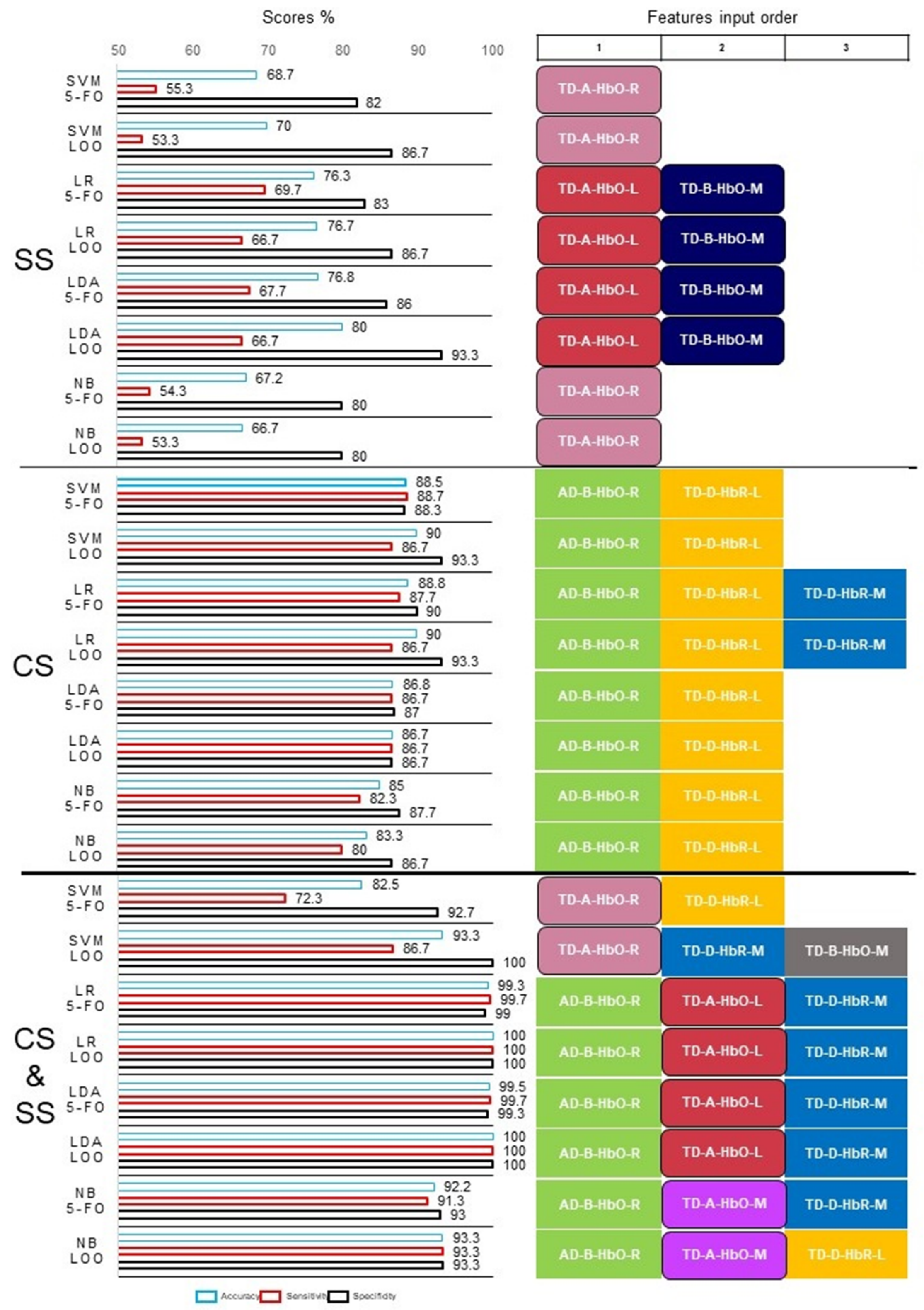


Figure 12. Performance scores achieved by each classification model using the subset of wrapper-selected features for shallow-signals (SS), clean-signals (CS) and SS + CS. Colored rectangles on the right identify the chosen features and their input order. Each specific feature has a different color and its label is the combination 'Group-Band-Chromophore-ROI', being TD = TD group, AD = ADHD, L = left, M = mid and R = right.

band *B* in which a prominent ISS peak can be seen (figure 9). TD group provides the next best feature in form of consistent HbO fluctuations at task frequency in the SS data (see the ISS peak in sub-band *A*). Finally, a very-slow CS-HbR component of TD group optimizes the classification (see the peak in *D*). The *shapelets* that generated these features are respectively identified by green, red and blue boxes in figures 10 and 11. Noteworthy, NB also shared the first feature while the second and third come from the same group, sub-band and chromophore but from an adjacent ROI.

Table 3 summarizes the results obtained by *wrappers* with SS + CS features. In all cases (except for SVM in sensitivity with five-fold) permutation testing indicated significance at $p < .001$. LR and LDA showed the highest scores in all metrics and also the narrowest CIs. Although highly significant, NB showed lower performance and larger ICs, whereas SVM performed the worst.

Noteworthy, when the full feature set (48 SS + 48 CS) was used to feed the *wrappers* for LR and LDA, we got exactly the same feature sub-set for both CV schemes and, hence, the same scores and statistics. Therefore, feature pre-selection did not lead to more optimistic solutions with inflated performance, albeit it did reduce computational cost.

4. Discussion

The present study aimed to assess the ability of a rhythmic mental math task to induce/modulate fNIRS oscillatory patterns (referred to here as *shapelets*) maximally representative of the ADHD or TD condition, and the feasibility of using them in automated classification of individuals. For this purpose, distinctive *shapelets* were first located on the basis of group synchronization strength at certain frequencies, and then simple measures of similarity were computed per-subject as features to train four popular machine learning algorithms suitable for binary classification. We found that with proper feature selection, classifiers achieved truly high predictive statistics when defining class membership from the individual oscillatory components. Noteworthy characteristics of the study are that it is based on a unimodal approach that focuses on data drawn exclusively from the fNIRS domain, uses Euclidean distance as the only type of feature to build the classification models, and features are linked to visually identifiable waveforms. In addition, we have limited ourselves to ready-to-use, inherently interpretable supervised linear algorithms (as available in MATLAB) without hyperparameter optimization. Therefore, we did not seek high performance at the expense of combining a variety of feature types and/or tuning classifiers, but rather explore the feasibility of proposing a methodological framework as a starting point for identifying hemodynamic biomarkers, in a way accessible and interpretable enough as to inform clinical practice.

4.1. Feature selection and classifiers performance

We corroborated that efficient feature selection greatly improves/stabilizes correct classification, being particularly important when the number of features exceeds the number of observations, as is often the case with small sample sizes [102]. Feature selection aims to discard irrelevant/redundant predictors, which improves predictive ability by reducing overfitting, leads to simpler classification models with less computational cost, and makes models easier to understand by knowing the most important variables. Since finding the optimal subset of features by exhaustively searching among all possible ones is often impractical, suboptimal selection methods are commonly used as workarounds, even with the awareness that they might not be optimal. Numerous studies have proposed different strategies to select features, however, no single method has been found that works best in all scenarios [99, 112]. Here, we employed a combined approach that consists of first obtaining a candidate set of features using a simple univariate statistical test (i.e. each feature is scored independently) to select the top ranked ones, and then applying a wrapper method to find the best subset among those features using the classification algorithm itself as evaluator. Given that the available data was relatively small and that complex search methods might be more prone to overfitting, we settled on a simpler strategy fully independent of any learning method [113]. Thus, for the first filtering step, we ranked the features by *F*-stat and then those with $p < .01$ were chosen. Despite its simplicity, the filter was good enough to achieve statistically significant accuracies (based on the binomial law at $p < 10^{-3}$) for all classifiers and CV schemes when CS and SS + CS features were used. In fact, with SS + CS the accuracies ranged between 83.3% and 90% depending on the case, which are comparable (some even better) to the previously reported values in [19–21]. Therefore, the features seemed to be intrinsically distinctive and worked well regardless of the different algorithmic architecture of the classifiers and CV partitioning.

Despite these promising results, some irrelevant features could potentially degrade performance and add unnecessary complexity. We addressed such a possibility by applying a *wrapper* to fine-tune feature selection. Specifically, we use an SFFS algorithm that is supposed to overcome simple sequential-selection by controlling the ‘nesting effect’ when, once selected, a feature cannot be dropped [99, 107]. In fact, we checked that our SFFS algorithm successfully conducted the backward steps to find better solutions. Remarkably, *wrappers* yielded very simple models with no more than 3 features, meaning a feature-to-sample ratio of 3/30 which is very unlikely to lead to overfitting [114]. Performance was particularly high when LR and LDA were trained on SS + CS features, achieving scores of around 100% on all metrics and CVs, both classifiers pointing to the same feature subset. While showing slightly lower scores, albeit >91%, NB also agreed in the same first feature and the other two shared similar properties as LR and LDA (figure 12). These results suggest that selected features

Table 3. Performance scores achieved by each classification model using the best subset of wrapper-selected features obtained from shallow- (SS) plus clean-signals (CS). Statistical significance is indicated by *p*-values and performance 95% CIs are represented within square brackets.

		SVM		LR		LDA		NB	
Signal	P. metric	5-FO	LOO	5-FO	LOO	5-FO	LOO	5-FO	LOO
CS & SS	Accuracy	82.5 $p < .001$ [68.2, 95.7]	93.3 $p < .001$ [80.0, 100]	99.3 $p < .001$ [95.8, 100]	100 $p < .001$ [96.7, 100]	99.5 $p < .001$ [96.7, 100]	100 $p < .001$ [96.7, 100]	92.2 $p < .001$ [80.3, 98.0]	93.3 $p < .001$ [83.3, 100]
	Sensitivity	72.3 $p = .046$ [56.0, 73.3]	86.7 $p < .001$ [66.7, 100]	99.7 $p < .001$ [93.3, 100]	100 $p < .001$ [93.3, 100]	99.7 $p < .001$ [92.0, 100]	100 $p < .001$ [93.3, 100]	91.3 $p < .001$ [73.3, 99.7]	93.3 $p < .001$ [80.0, 100]
	Specificity	92.7 $p < .001$ [78.2, 100]	100 $p < .001$ [86.7, 100]	99.0 $p < .001$ [91.8, 100]	100 $p < .001$ [93.3, 100]	99.3 $p < .001$ [90.4, 100]	100 $p < .001$ [93.3, 100]	93.0 $p < .001$ [80, 100]	93.3 $p < .001$ [80.0, 100]
	Features	2	3	3	3	3	3	3	3

Note: LOO = leave-one-out; 5-FO = 5-fold; SVM = support vector machine; LR = logistic regression; LDA = linear discriminant analysis; NB = naïve Bayes.

are quite stable and survive different classification algorithms and CV schemes. Such stability can also be seen in the CS case, in which an even wider consensus is shared across classifiers (figure 12). Therefore, it appears that the observed performance was not due to an over-selection effect [115], but to the true distinctive nature of the features.

Surprisingly, SVM performed the worst in the SS + CS case, particularly with the five-fold CV where it achieved the lowest scores and also the widest CIs. In contrast, the other classifiers were not affected by the CV, reaching similar scores and CIs with both schemes (table 3). Most likely, SVM performance degraded due to lack of proper regularization, leading to insufficient penalty for misclassification. As mentioned, one of the goals was to avoid any hyperparameter tuning and therefore we preferred to rule out SVM in the context of the present study.

As can be seen from figure 12 and table 3, LR and LDA won in the classification task when trained with SS + CS features. They both agreed on the same *wrapper* solution, which was exactly the same for the features preselected by filtering and for the full set. Also, they achieved comparable scores (>99%) in all metrics (significant at $p < .001$, but now using permutation tests to define the null hypothesis). In addition, the 95% CIs were similarly narrow, although skewed due to scores being close to the 100% ceiling. CV is known to tend towards narrower confidence bounds as accuracy approaches 100% but increasingly wide and asymmetric as sample size decreases, which can lead to under-estimate prediction errors and specially with LOO CV [116]. Despite having only 30 samples, we found that the lower limit of the CI for accuracy was never <95.8% and was less than 5% away from the mean in all cases, a deviation below the overall 15% expected for a binary classification with this number of samples [116]. The results suggest that the few selected features meet the statistical assumptions required by LR and LDA to make observations highly separable into 2-classes. The rest of this discussion focuses on LR and LDA and their three wrapper-selected shared features.

4.2. CV agreement

As reported in [12], LOO and *k*-fold are the most commonly adopted CV methods in ADHD studies; plus hold-out which seems more appropriate for large datasets. Some researchers have suggested that LOO may be more useful in a diagnostic scenario [98, 117], whereas others recommend *k*-fold or repeated random splits for more stable estimates [97, 118]. In either case, it is known that CV is compromised by small sample sizes, particularly if there are many predictors, which tend to overestimate predictive accuracy to a variable degree depending on the particularities of the study [14, 119]. In light of this, we tested LOO and five-fold expecting differences in performance, with LOO showing more optimistic scores and larger confidence bounds [116]. However, we found that both gave very similar results with LR and LDA (table 3), suggesting that the three chosen features are good enough predictors to yield stable CV measures regardless of method.

4.3. Contribution of SS and CS to classification

The present work did not rely on activation/deactivation measures as usual in fNIRS studies, but rather on the distinctive information content of frequency components. Among others, this is one of the reasons why we cautiously prefer to consider CS as clean/corrected DSs, without further assumptions about their true nature. On another hand, we did not consider SSs as nuisance components to be discarded, but as potential carriers of representative information due to the close interaction between cognitive and autonomic functioning [120–122]. In fact, we got accuracies close to 100% when SS and CS features were combined, which out-perform other classification studies using unimodal fNIRS data.

Remarkably, the first-selected feature came from CS of the ADHD group in form of a consistent HbO pattern in the *B* band (7.7–20.3 mHz) of right-ROI (figures 11 and 12). Notably, this pattern increasingly synchronizes across participants as mental task progresses and peaks near the end, while TD group peaks faster and decreases towards the end (similar time-courses can be seen in all three ROIs). It seems that some kind of underlying hemodynamic activity works differently in ADHD at that frequency. Resting-state fMRI (rs-fMRI) studies have reported frequency-specific ADHD abnormalities in local spontaneous activity of multiple brain regions, detected not only in the conventional fMRI range of 10–80 mHz [29], but also in narrower sub-bands such as 0–10, 10–27 and 27–73 mHz among others [31]. However, probably due to the high heterogeneity of ADHD presentation, consistent conclusions have not yet been reached [32]. Nevertheless, together with the findings from rs-fMRI connectivity studies [123, 124], evidence suggests that low-frequency fluctuations analysis could provide valuable insights into the dysfunction of brain-networks that has been observed in many ADHD studies (for reviews, see [7, 125]). As can be deduced from the studies mentioned, few agreed conclusions can be drawn from the large number of heterogeneous reports. Nevertheless, reports frequently concur on the key role of the PFC in the altered relationships between DMN and attention/salience networks observed in ADHD, likely related to a delayed cortical maturation [126, 127]. Since our NIRS probe interrogates part of the PFC, there is a possibility that this distinctive ADHD pattern we found is due to delayed/interfered DMN deactivation during the task, which is reflected as a specific frequency component. In line with this finding, Salmi *et al* [128] found increased impulsivity-associated synchronization in the medial PFC of ADHD patients during a naturalistic attention task.

Similar discussion could be applied to the third feature that also arose from CS, but from an HbR *shapelet* in the ultra-low *D*-band of TD group (also visible in all ROIs). It can be seen a consistent pattern of increasing at task onset, slowly returning to lower levels and raising again at the end (figure 11), which is absent in the ADHD group. Whether the initial effect is due to sustained oxygen consumption, deactivation, or both being interrelated, is inherently difficult to elucidate with fNIRS data alone. Interestingly, it has been reported that ultra-slow BOLD signals (0-10 mHz) seem to better differentiate ADHD from TD [31]. While more research is needed, the importance of assessing both HbO and HbR should not be underestimated.

The second-selected feature arose from an SS *shapelet* of HbO in the task frequency *A*-band and contributed significantly to improving classification (figure 12). Obviously, this oscillatory pattern is not of cerebral origin but rather reflects a coordinated systemic (even local) hemodynamic activity recorded in superficial tissue layers. Specifically, this SS feature pertains to TD boys who show greater within-group synchronization than the ADHD group, visible also in the other two ROIs (figure 10). This finding is in line with other studies using thermal imaging [129] or EKG entropy measures [130] to differentiate ADHD with promising accuracy. This lack of synchrony probably reflects the inability of ADHD boys to coordinate autonomic resources with cognitive demands [131].

4.4. Accuracy, sensitivity and specificity

We instructed the *wrappers* to favor specificity first and accuracy second. The rationale behind is that, by favoring specificity, classifiers are less prone to false positives and so more reliable when predicting ADHD. Marking a child as ADHD when it is not true can lead to risky drug treatments, social stigma and other annoyances. Conversely, without major drawbacks, a false negative can be easily ruled-out through a complementary evaluation and/or follow-up over time. However, since the specificity scored so high and the sample was exactly balanced, the accuracy (and sensitivity) rose correspondingly, with very little difference between the metrics. Although useless in the present work, we believe that defining the order of importance of the metrics could lead to a better selection of relevant data when considering the goals and particularities of clinical trials. On another hand, albeit not predictive measures (prevalence was not considered), the metrics used here suffice as a proof-of-concept, as they are considered useful for evaluating screening/diagnostic tests in clinical research [132, 133].

4.5. Limitations and future work

The current study is not intended to validate an ADHD diagnostic tool for use in a clinical setting, but to propose a methodological framework as a proof-of-concept to find functional biomarkers to help develop such a tool. However, some issues need to be addressed in order to properly frame the current findings. First, despite the high scores achieved, the limited sample size could have led to overestimating the performance of the classification. However, since ours is a fairly pure sample of only boys within a narrow age range and focused on combined ADHD subtype, the effective sample size is comparable or greater than those of other classification studies. The homogeneity of our sample can be seen as a limitation that would prevent generalizing the findings to other groups, but we feel that it is rather a strength of the study. ADHD condition is so heterogeneous that trying to find an all-in-one solution can be a difficult goal to achieve. We believe that tailoring functional biomarkers for specific use-case samples (e.g. women, drug-naïve children, adolescents, etc) may be more productive and easily implementable as a decision tree. Obviously, the approach proposed here can be easily applied to those other use-cases, work that we currently have under way. Whether this approach can benefit from extended multi-distance measurements with partial path length correction, as proposed in [134], will require future research.

Second, the unavailability of a separate dataset for validation may cast doubt on the generalizability of the models. Therefore, to test the usefulness of the current method in clinical decision support, further research with a larger sample size and external validation will be necessary.

5. Conclusion

Over the last decade, considerable efforts have been made in the development of AI-based models to aid clinical decisions in brain disorders such as ADHD, but with little impact on the healthcare workflow. Despite promising results in the field of neuroimaging, substantial problems remain that make it difficult to bridge the gap between research and daily clinic. Among other challenges, finding reliable biomarkers via technically feasible and explainable methods is essential to enable AI implementations as trustworthy clinical decision support systems.

To contribute to the objective assessment of ADHD, we provide a proof-of-concept that very-low frequency fNIRS fluctuations induced by a rhythmic mental task accurately differentiate ADHD boys (average age 11.9 years) from non-ADHD controls at an individual level. We propose a method, based on synchronization analysis in the time-frequency plane, to find distinctive oscillatory patterns from which to extract simple distance-based features following a *shapelet*-like approach. We also propose an SFFS *wrapper* algorithm combined with ML linear models as evaluators for efficient feature selection. The results showed that with only three key features, LR and LDA classifiers achieved accuracy, sensitivity and specificity scores of nearly 100%, outperforming other fNIRS studies. An advantage of this proposal is greater transparency of the classification outcomes, since LR and LDA models are linearly interpretable and key features are visualizable as physiologically significant hemodynamic patterns. We also suggest that predictive models can be improved by targeting specific samples of ADHD patients. Our observation of altered specific frequency components is compatible with previous studies pointing to ADHD-associated abnormalities in cortical maturation and neural networks connectivity, particularly those related to PFC functioning and autonomic-cognitive interaction. To our knowledge, the methodological approach presented here has not been previously tested in fNIRS-based ADHD studies. It could also be used to assess other types of brain disorders. Despite our promising results, we emphasize that further rigorous validation is required to confirm the ability of the method to provide robust biomarkers as adjunct indicators applicable into the clinical practice.

Data availability statement

The data generated and/or analyzed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgments

The authors are immensely grateful to the participants and their families who selflessly participated in this research. We acknowledge the technical assistance of Mr Miguel Ibañez and Mr Alejandro Mendez.

Conflict of interest

Joaquín Ibañez-Ballesteros and Carlos Belmonte report that they are inventors of patents licensed to Newmanbrain, S L and co-founders and scientific advisors of Newmanbrain S L, the company responsible of manufacturing the NIRS device used in this research.

Ethical statement

All procedures performed in this study involving human participants were approved by the Ethical Committee of Universidad Miguel Hernandez and conform to the 1964 Helsinki declaration and its later amendments, as well as with local regulations. Written informed consent was obtained from parents and verbal assent from all the boys.

Funding

This work was conducted in absence of any funding.

CRediT authorship contribution statement

Sergio Ortuño-Miró: Conceptualization, Methodology, Formal analysis, Investigation, Writing—Original Draft. Sergio Molina-Rodríguez: Conceptualization, Methodology, Formal analysis, Investigation, Writing—Original Draft. Carlos Belmonte: Conceptualization, Writing—Review & Editing. Joaquín Ibañez-Ballesteros: Conceptualization, Methodology, Formal analysis, Software, Writing—Review & Editing, Supervision. All authors have reviewed and approved the manuscript.

ORCID iD

Joaquín Ibañez-Ballesteros https://orcid.org/0000-0001-8606-4221

References

[1] Biederman J 2005 Attention-deficit/hyperactivity disorder: a selective overview *Biol. Psychiatry* **57** 1215–20

[2] Canals J, Morales-Hidalgo P, Jané M C and Domènech E 2018 ADHD prevalence in Spanish preschoolers: comorbidity, socio-demographic factors, and functional consequences *J. Atten. Disord.* **22** 143–53

[3] Thapar A and Cooper M 2016 Attention deficit hyperactivity disorder *Lancet* **387** 1240–50

[4] Gadow K D, Drabick D A G, Loney J, Sprafkin J, Salisbury H, Azizian A and Schwartz J 2004 Comparison of ADHD symptom subtypes as source-specific syndromes *J. Child Psychol. Psychiatry* **45** 1135–49

[5] Campbell S B 2000 Attention-deficit/hyperactivity disorder: a developmental view *Handbook of Developmental Psychopathology* (Boston, MA: Springer) pp 383–401

[6] Faraone S V *et al* 2021 The world federation of ADHD international consensus statement: 208 evidence-based conclusions about the disorder *Neurosci. Biobehav. Rev.* **128** 789–818

[7] Posner J, Polanczyk G V and Sonuga-Barke E 2020 Attention-deficit hyperactivity disorder *Lancet* **395** 450–62

[8] American Psychiatric Association 2013 *Diagnostic and Statistical Manual of Mental Disorders* (Arlington, VA: American Psychiatric Association)

[9] Adesman A R 2001 The diagnosis and management of attention-deficit/hyperactivity disorder in pediatric patients *Prim. Care Companion J. Clin. Psychiatry* **3** 66–77

[10] Halperin J M, Bédard A C V and Curchack-Lichtin J T 2012 Preventive interventions for ADHD: a neurodevelopmental perspective *Neurotherapeutics* **9** 531–41

[11] Pereira-Sanchez V and Castellanos F X 2021 Neuroimaging in attention-deficit/hyperactivity disorder *Curr. Opin. Psychiatry* **34** 105–11

[12] Loh H W, Ooi C P, Barua P D, Palmer E E, Molinari F and Acharya U R 2022 Automated detection of ADHD: current trends and future perspective *Comput. Biol. Med.* **146** 105525

[13] ADHD-200-Consortium 2012 The ADHD-200 Consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience *Front. Syst. Neurosci.* **6** 62

[14] Arbabshirani M R, Plis S, Sui J and Calhoun V D 2017 Single subject prediction of brain disorders in neuroimaging: promises and pitfalls *Neuroimage* **145** 137–65

[15] Pulini A A, Kerr W T, Loo S K and Lenartowicz A 2019 Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4** 108–20

[16] Mauri M, Nobile M, Bellina M, Crippa A and Brambilla P 2018 Light up ADHD: I. Cortical hemodynamic responses measured by functional near infrared spectroscopy (fNIRS) *J. Affect. Disord.* **234** 358–64

[17] Pinti P, Tachtsidis I, Hamilton A, Hirsch J, Aichelburg C, Gilbert S and Burgess P W 2018 The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience *Ann. New York Acad. Sci.* **1464** 1–25

[18] Güven A, Altınkaynak M, Dolu N, İzzetoğlu M, Pektaş F, Özmen S, Demirci E and Batbat T 2020 Combining functional near-infrared spectroscopy and EEG measurements for the diagnosis of attention-deficit hyperactivity disorder *Neural Comput. Appl.* **32** 8367–80

[19] Monden Y *et al* 2015 Individual classification of ADHD children by right prefrontal hemodynamic responses during a go/no-go task as assessed by fNIRS *NeuroImage Clin.* **9** 1–12

[20] Crippa A, Salvatore C, Molteni E, Mauri M, Salandi A, Trabattoni S, Agostoni C, Molteni M, Nobile M and Castiglioni I 2017 The utility of a computerized algorithm based on a multi-domain profile of measures for the diagnosis of attention deficit/hyperactivity disorder *Front. Psychiatry* **8** 1–10

[21] Gu Y, Miao S, Han J, Liang Z, Ouyang G, Yang J and Li X 2018 Identifying ADHD children using hemodynamic responses during a working memory task measured by functional near-infrared spectroscopy *J. Neural Eng.* **15** 035005

[22] Takahashi T, Takikawa Y, Kawagoe R, Shibuya S, Iwano T and Kitazawa S 2011 Influence of skin blood flow on near-infrared spectroscopy signals measured on the forehead during a verbal fluency task *Neuroimage* **57** 991–1002

[23] Haeussinger F B, Dresler T, Heinzel S, Schecklmann M, Fallgatter A J and Ehlis A-C 2014 Reconstructing functional near-infrared spectroscopy (fNIRS) signals impaired by extra-cranial confounds: an easy-to-use filter method *Neuroimage* **95** 69–79

[24] Tachtsidis I and Scholkmann F 2016 False positives and false negatives in functional near-infrared spectroscopy: issues, challenges, and the way forward *Neurophotonics* **3** 031405

[25] Caldwell M, Scholkmann F, Wolf U, Wolf M, Elwell C and Tachtsidis I 2016 Modelling confounding effects from extracerebral contamination and systemic factors on functional near-infrared spectroscopy *Neuroimage* **143** 91–105

[26] Kirilina E, Jelzow A, Heine A, Niessing M, Wabnitz H, Brühl R, Ittermann B, Jacobs A M and Tachtsidis I 2012 The physiological origin of task-evoked systemic artefacts in functional near infrared spectroscopy *Neuroimage* **61** 70–81

[27] Molina-Rodríguez S, Mirete-Fructuoso M, Martínez L and Ibañez-Ballesteros J 2022 Frequency-domain analysis of fNIRS fluctuations induced by rhythmic mental arithmetic *Psychophysiology* **1** 1–25

[28] Yu-Feng W, Yong H, Chao-Zhe Z, Qing-Jiu C, Man-Qiu S, Meng L, Li-Xia T, Tian-Zi J and Yu-Feng W 2007 Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI *Brain Dev.* **29** 83–91

[29] Tang C, Wei Y, Zhao J and Nie J 2018 Different developmental pattern of brain activities in ADHD: a study of resting-state fMRI *Dev. Neurosci.* **40** 246–57

[30] Zhang J, Kendrick K M, Lu G and Feng J 2015 The fault lies on the other side: altered brain functional connectivity in psychiatric disorders is mainly caused by counterpart regions in the opposite hemisphere *Cereb. Cortex* **25** 3475–86

[31] Yu X, Yuan B, Cao Q, An L, Wang P, Vance A, Silk T J, Zang Y, Wang Y and Sun L 2016 Frequency-specific abnormalities in regional homogeneity among children with attention deficit hyperactivity disorder: a resting-state fMRI study *Sci. Bull.* **61** 682–92

[32] Wang J-B, Zheng L-J, Cao Q-J, Wang Y-F, Sun L, Zang Y-F and Zhang H 2017 Inconsistency in abnormal brain activity across cohorts of adhd-200 in children with attention deficit hyperactivity disorder *Front. Neurosci.* **11** 1–10

[33] Scholkmann F, Gerber U, Wolf M and Wolf U 2013 End-tidal CO$_2$: an important parameter for a correct interpretation in functional brain studies using speech tasks *Neuroimage* **66** 71–79

[34] Zimeo Morais G A, Scholkmann F, Balardin J B, Furucho R A, de Paula R C V, Biazoli C E and Sato J R 2017 Non-neuronal evoked and spontaneous hemodynamic changes in the anterior temporal region of the human head may lead to misinterpretations of functional near-infrared spectroscopy signals *Neurophotonics* **5** 1

[35] Orihuela-Espina F, Leff D R, James D R C, Darzi A W and Yang G Z 2010 Quality control and assurance in functional near infrared spectroscopy (fNIRS) experimentation *Phys. Med. Biol.* **55** 3701–24

[36] Kocsis L, Herman P and Eke A 2006 The modified Beer-Lambert law revisited *Phys. Med. Biol.* **51** N91

[37] Delpy D T, Cope M, Van Der Zee P, Arridge S, Wray S and Wyatt J 1988 Estimation of optical pathlength through

tissue from direct time of flight measurement *Phys. Med. Biol.* **33** 1433–42

[38] Huppert T J, Diamond S G, Franceschini M A and Boas D A 2009 HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain *Appl. Opt.* **48** 0–33

[39] Reddy P, Izzetoglu M, Shewokis P A, Sangobowale M, Diaz-Arrastia R and Izzetoglu K 2021 Evaluation of fNIRS signal components elicited by cognitive and hypercapnic stimuli *Sci. Rep.* **11** 23457

[40] Scholkmann F, Kleiser S, Metz A J, Zimmermann R, Mata Pavia J, Wolf U and Wolf M 2014 A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology *Neuroimage* **85** 6–27

[41] Brigadoi S and Cooper R J 2015 How short is short? Optimum source–detector distance for short-separation channels in functional near-infrared spectroscopy *Neurophotonics* **2** 025005

[42] Gagnon L, Cooper R J, Yücel M A, Perdue K L, Greve D N and Boas D A 2012 Short separation channel location impacts the performance of short channel regression in NIRS *Neuroimage* **59** 2518–28

[43] Yücel M A, Selb J J, Huppert T J, Franceschini M A and Boas D A 2017 Functional near infrared spectroscopy: enabling routine functional brain imaging *Curr. Opin. Biomed. Eng.* **4** 78–86

[44] Gagnon L, Yücel M A, Boas D A and Cooper R J 2014 Further improvement in reducing superficial contamination in NIRS using double short separation measurements *Neuroimage* **85** 127–35

[45] Holland P W and Welsch R E 1977 Robust regression using iteratively reweighted least-squares *Commun. Stat. Theory Methods* **6** 813–27

[46] Tak S, Uga M, Flandin G, Dan I and Penny W D 2016 Sensor space group analysis for fNIRS data *J. Neurosci. Methods* **264** 103–12

[47] Plichta M M, Herrmann M J, Baehne C G, Ehlis A-C, Richter M M, Pauli P and Fallgatter A J 2006 Event-related functional near-infrared spectroscopy (fNIRS): are the measurements reliable? *Neuroimage* **31** 116–24

[48] Schecklmann M, Ehlis A-C, Plichta M M and Fallgatter A J 2008 Functional near-infrared spectroscopy: a long-term reliable tool for measuring brain activity during verbal fluency *Neuroimage* **43** 147–55

[49] Näsi T, Mäki H, Hiltunen P, Heiskala J, Nissilä I, Kotilahti K and Ilmoniemi R J 2013 Effect of task-related extracerebral circulation on diffuse optical tomography: experimental data and simulations on the forehead *Biomed. Opt. Express* **4** 412

[50] Nambu I, Ozawa T, Sato T, Aihara T, Fujiwara Y, Otaka Y, Osu R, Izawa J and Wada Y 2017 Transient increase in systemic interferences in the superficial layer and its influence on event-related motor tasks: a functional near-infrared spectroscopy study *J. Biomed. Opt.* **22** 035008

[51] Welch P 1967 The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms *IEEE Trans. Audio Electroacoust.* **15** 70–73

[52] Ilvedson C R 1998 *Transfer Function Estimates Using Time-Frequency Analysis* (Cambridge: Massachusetts Institute of Technology)

[53] Aarabi A and Huppert T J 2016 Characterization of the relative contributions from systemic physiological noise to whole-brain resting-state functional near-infrared spectroscopy data using single-channel independent component analysis *Neurophotonics* **3** 025004

[54] Maris E and Oostenveld R 2007 Nonparametric statistical testing of EEG- and MEG-data *J. Neurosci. Methods* **164** 177–90

[55] Cohen L 1995 *Time-frequency Analysis* (Englewood Cliffs, NJ: Prentice Hall)

[56] Gröchenig K 2001 *Foundations of Time-Frequency Analysis* (Boston, MA: Birkhäuser Boston)

[57] Morales S and Bowers M E 2022 Time-frequency analysis methods and their application in developmental EEG data *Dev. Cogn. Neurosci.* **54** 101067

[58] Glerean E, Salmi J, Lahnakoski J M, Jääskeläinen I P and Sams M 2012 Functional magnetic resonance imaging phase synchronization as a measure of dynamic functional connectivity *Brain Connect.* **2** 91–101

[59] Wacker M and Witte H 2013 Time-frequency techniques in biomedical signal analysis: a tutorial review of similarities and differences *Methods Inf. Med.* **52** 279–96

[60] Munia T T K and Aviyente S 2019 Time-frequency based phase-amplitude coupling measure for neuronal oscillations *Sci. Rep.* **9** 1–15

[61] Liu T, Luo Z, Huang J and Yan S 2018 A comparative study of four kinds of adaptive decomposition algorithms and their applications *Sensors* **18** 1–51

[62] Kijewski-Correa T and Kareem A 2006 Efficacy of Hilbert and wavelet transforms for time-frequency analysis *J. Eng. Mech.* **132** 1037–49

[63] Cohen A and Kovacevic J 1996 Wavelets: the mathematical background *Proc. IEEE* **84** 514–22

[64] Hramov A E, Koronovskii A A, Makarov V A, Pavlov A N and Sitnikova E 2015 *Wavelets in Neuroscience* (Berlin: Springer)

[65] Sadowsky J 1996 Investigation of signal characteristics using the continuous wavelet transform *Johns Hopkins APL Tech. Dig.* **17** 258–69

[66] Olhede S C and Walden A T 2002 Generalized Morse wavelets *IEEE Trans. Signal Process.* **50** 2661–70

[67] Lilly J M and Olhede S C 2010 On the analytic wavelet transform *IEEE Trans. Inf. Theory* **56** 4135–56

[68] Lilly J M and Olhede S C 2012 Generalized Morse wavelets as a superfamily of analytic wavelets *IEEE Trans. Signal Process.* **60** 6036–41

[69] Suwansawang S and Halliday D M 2017 Time-frequency based coherence and phase locking value analysis of human locomotion data using generalized Morse wavelets *BIOSIGNALS 2017-10th Int. Conf. Bio-Inspired Syst. Signal Process* vol 4 pp 34–41

[70] Wachowiak M P, Wachowiak-Smolíková R, Johnson M J, Hay D C, Power K E and Williams-Bell F M 2018 Quantitative feature analysis of continuous analytic wavelet transforms of electrocardiography and electromyography *Phil. Trans. R. Soc.* A **376** 20170250

[71] Wiklendt L, Brookes S J H, Costa M, Travis L, Spencer N J and Dinning P G 2020 A novel method for electrophysiological analysis of EMG signals using mesaclip *Front. Physiol.* **11** 1–10

[72] Nakhnikian A, Ito S, Dwiel L L, Grasse L M, Rebec G V, Lauridsen L N and Beggs J M 2016 A novel cross-frequency coupling detection method using the generalized Morse wavelets *J. Neurosci. Methods* **269** 61–73

[73] Huggins C J, Escudero J, Parra M A, Scally B, Anghinah R, Vitória Lacerda De Araújo A, Basile L F and Abasolo D 2021 Deep learning of resting-state electroencephalogram signals for three-class classification of Alzheimer's disease, mild cognitive impairment and healthy ageing *J. Neural Eng.* **18** 046087

[74] Ajith K, Menaka R and Kumar S S 2021 EEG based mental state analysis *J. Phys.: Conf. Ser.* **1911** 012014

[75] Perpetuini D, Cardone D, Filippini C, Chiarelli A M and Merla A 2021 A motion artifact correction procedure for fNIRS signals based on wavelet transform and infrared thermography video tracking *Sensors* **21** 5117

[76] Lilly J M 2017 Element analysis: a wavelet-based method for analysing time-localized events in noisy time series *Proc. R. Soc.* A **473** 20160776

[77] Nastase S A, Gazzola V, Hasson U and Keysers C 2019 Measuring shared responses across subjects using intersubject correlation *Soc. Cogn. Affect. Neurosci.* **14** 669–87

[78] Hasson U, Nir Y, Levy I, Fuhrmann G and Malach R 2004 Intersubject synchronization of cortical activity during natural vision *Science* **303** 1634–40

[79] Bolt T, Nomi J S, Vij S G, Chang C and Uddin L Q 2018 Inter-subject phase synchronization for exploratory analysis of task-fMRI *Neuroimage* **176** 477–88

[80] Kauppi J-P, Pajula J and Tohka J 2014 A versatile software package for inter-subject correlation based analyses of fMRI *Front. Neuroinform.* **8** 1–13

[81] Berens P 2009 CircStat: a MATLAB toolbox for circular statistics *J. Stat. Softw.* **31** 1–21

[82] Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Methods* **134** 9–21

[83] Cohen M X 2014 *Analyzing Neural Time Series Data* (Cambridge, MA: MIT Press)

[84] Tallon-Baudry C, Bertrand O, Delpuech C and Pernier J 1996 Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human *J. Neurosci.* **16** 4240–9

[85] Bastos A M and Schoffelen J-M 2016 A tutorial review of functional connectivity analysis methods and their interpretational pitfalls *Front. Syst. Neurosci.* **9** 1–23

[86] Richter C G, Thompson W H, Bosman C A and Fries P 2015 A jackknife approach to quantifying single-trial correlation between covariance-based metrics undefined on a single-trial basis *Neuroimage* **114** 57–70

[87] Fulcher B D 2018 Feature-based time-series analysis *Feature Engineering for Machine Learning and Data Analytics* Dong G and Liu H (Boca Raton, FL: CRC press) pp 87–116

[88] Ye L and Keogh E 2011 Time series shapelets: a novel technique that allows accurate, interpretable and fast classification *Data Min. Knowl. Discov.* **22** 149–82

[89] Hills J, Lines J, Baranauskas E, Mapp J and Bagnall A 2014 Classification of time series by shapelet transformation *Data Min. Knowl. Discov.* **28** 851–81

[90] Richard A, Mayag B, Talbot F, Tsoukias A and Meinard Y 2020 Transparency of Classification Systems for Clinical Decision Support *Information Processing and Management of Uncertainty in Knowledge-Based Systems* 1239 (London: Nature Publishing Group) pp 99–113

[91] Hearst M A, Dumais S T, Osuna E, Platt J and Scholkopf B 1998 Support vector machines *IEEE Intell. Syst. Appl.* **13** 18–28

[92] Hosmer D W Jr, Lemeshow S and Sturdivant R X 2013 *Applied Logistic Regression* vol 398 (New York: Wiley)

[93] Fisher R A 1936 The use of multiple measurements in taxonomic problems *Ann. Eugen.* **7** 179–88

[94] Kuncheva L I 2006 On the optimality of Naïve Bayes with dependent binary features *Pattern Recognit. Lett.* **27** 830–7

[95] Hand D J and Yu K 2001 Idiot's bayes: not so stupid after all? *Int. Stat. Rev.* **69** 385

[96] Pereira F, Mitchell T and Botvinick M 2009 Machine learning classifiers and fMRI: a tutorial overview *Neuroimage* **45** S199–209

[97] Hastie T, Tibshirani R, Friedman J H and Friedman J H 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* vol 2 (New York: Springer)

[98] Saeb S, Lonini L, Jayaraman A, Mohr D C and Kording K P 2017 The need to approximate the use-case in clinical machine learning *Gigascience* **6** 1–9

[99] Somol P, Novovičová J and Pudil P 2010 Efficient feature subset selection and subset size optimization *Pattern Recognition Recent Advances* ed A Herout (Rijeka: Intech) (https://doi.org/10.5772/224)

[100] Liu H and Yu L 2005 Toward integrating feature selection algorithms for classification and clustering *IEEE Trans. Knowl. Data Eng.* **17** 491–502

[101] Amann J *et al* 2022 To explain or not to explain?—artificial intelligence explainability in clinical decision support systems ed H H-S Lu *PLoS Digit. Heal.* **1** e0000016

[102] Jovic A, Brkic K and Bogunovic N 2015 A review of feature selection methods with applications *2015 38th Int. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (IEEE) pp 1200–5

[103] Kim J-H 2009 Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap *Comput. Stat. Data Anal.* **53** 3735–45

[104] Luque A, Carrasco A, Martín A and de Las Heras A 2019 The impact of class imbalance in classification performance metrics based on the binary confusion matrix *Pattern Recognit.* **91** 216–31

[105] Hossin M and Sulaiman M N 2015 A review on evaluation metrics for data classification evaluations *Int. J. Data Min. Knowl. Manage. Process* **5** 01–11

[106] Combrisson E and Jerbi K 2015 Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy *J. Neurosci. Methods* **250** 126–36

[107] Pudil P, Novovičová J and Kittler J 1994 Floating search methods in feature selection *Pattern Recognit. Lett.* **15** 1119–25

[108] Bzdok D 2017 Classical statistics and statistical learning in imaging neuroscience *Front. Neurosci.* **11** 1–23

[109] Golland P and Fischl B R 2003 Permutation tests for classification: towards statistical significance in image-based studies *Information Processing in Medical Imaging* vol 18 (Berlin: Springer) pp 330–41

[110] Fu W J, Carroll R J and Wang S 2005 Estimating misclassification error with small samples via bootstrap cross-validation *Bioinformatics* **21** 1979–86

[111] Kriegeskorte N, Simmons W K, Bellgowan P S and Baker C I 2009 Circular analysis in systems neuroscience: the dangers of double dipping *Nat. Neurosci.* **12** 535–40

[112] Remeseiro B and Bolon-Canedo V 2019 A review of feature selection methods in medical applications *Comput. Biol. Med.* **112** 25–29

[113] Reunanen J 2003 Overfitting in making comparisons between variable selection methods *J. Mach. Learn. Res.* **3** 1371–82

[114] Vabalas A, Gowen E, Poliakoff E, Casson A J and Hernandez-Lemus E 2019 Machine learning algorithm validation with a limited sample size *PLoS One* **14** 1–20

[115] Raudys S 2006 Feature over-selection *Structural, Syntactic, and Statistical Pattern Recognition* (Berlin: Springer) pp 622–31

[116] Varoquaux G 2018 Cross-validation failure: small sample sizes lead to large error bars *Neuroimage* **180** 68–77

[117] Tougui I, Jilbab A and El Mhamdi J 2021 Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications *Healthc. Inform. Res.* **27** 189–99

[118] Varoquaux G, Raamana P R, Engemann D A, Hoyos-Idrobo A, Schwartz Y and Thirion B 2017 Assessing and tuning brain decoders: cross-validation, caveats, and guidelines *Neuroimage* **145** 166–79

[119] Flint C *et al* 2021 Systematic misestimation of machine learning performance in neuroimaging studies of depression *Neuropsychopharmacology* **46** 1510–7

[120] Forte G, De Pascalis V, Favieri F and Casagrande M 2019 Effects of blood pressure on cognitive performance: a systematic review *J. Clin. Med.* **9** 34

[121] Forte G, Favieri F and Casagrande M 2019 Heart rate variability and cognitive function: a systematic review *Front. Neurosci.* **13** 1–11

[122] Wang X, Liu B, Xie L, Yu X, Li M and Zhang J 2016 Cerebral and neural regulation of cardiovascular activity during mental stress *Biomed. Eng. Online* **15** 335–47

[123] Wang X-H and Li L 2015 Altered temporal features of intrinsic connectivity networks in boys with combined type of attention deficit hyperactivity disorder *Eur. J. Radiol.* **84** 947–54

[124] Castellanos F X and Proal E 2012 Large-scale brain systems in ADHD: beyond the prefrontal-striatal model *Trends Cogn. Sci.* **16** 17–26

[125] Castellanos F X and Aoki Y 2016 Intrinsic functional connectivity in attention-deficit/hyperactivity disorder: a science in development *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **1** 253–61

[126] Shaw P K, Eckstrand W, Sharp J, Blumenthal J P, Lerch D, Greenstein L, Clasen A and Evans J 2007 Attention-deficit/hyperactivity disorder is characterized by a delay in cortical maturation *Proc. Natl Acad. Sci.* **104** 19649–54

[127] Halperin J M and Schulz K P 2006 Revisiting the role of the prefrontal cortex in the pathophysiology of attention-deficit/hyperactivity disorder *Psychol. Bull.* **132** 560–81

[128] Salmi J, Metwaly M, Tohka J, Alho K, Leppämäki S, Tani P, Koski A, Vanderwal T and Laine M 2020 ADHD desynchronizes brain activity during watching a distracted multi-talker conversation *Neuroimage* **216** 116352

[129] Coben R and Myers T E 2009 Sensitivity and specificity of long wave infrared imaging for attention-deficit/ hyperactivity disorder *J. Atten. Disord.* **13** 56–65

[130] Koh J E W, Ooi C P, Lim-Ashworth N S, Vicnesh J, Tor H T, Lih O S, Tan R S, Acharya U R and Fung D S S 2022 Automated classification of attention deficit hyperactivity disorder and conduct disorder using entropy features with ECG signals *Comput. Biol. Med.* **140** 105120

[131] Bellato A, Arora I, Hollis C and Groom M J 2020 Is autonomic nervous system function atypical in attention deficit hyperactivity disorder (ADHD)? A systematic review of the evidence *Neurosci. Biobehav. Rev.* **108** 182–206

[132] Trevethan R 2017 Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice *Front. Public Health* **5** 1–7

[133] Eusebi P 2013 Diagnostic accuracy measures *Cerebrovasc. Dis.* **36** 267–72

[134] Song X, Chen X, Chen L, An X and Ming D 2020 Performance improvement for detecting brain function using fNIRS: a multi-distance probe configuration with PPL method *Front. Hum. Neurosci.* **14** 1–11