The blessing of Dimensionality: Feature Selection outperforms functional connectivity-based feature transformation to classify ADHD subjects from EEG patterns of phase synchronisation

Ernesto Pereda[1,2,3]☯*, Miguel García-Torres[4]☯, Belén Melián-Batista[5], Soledad Mañas[6], Leopoldo Méndez[6], Julián J. González[7]

1 Electrical Engineering and Bioengineering Group, Department of Industrial Engineering & Instituto Universitario de Neurociencia (IUNE), Universidad de La Laguna, Santa Cruz de Tenerife, Spain, **2** Lab. of Cognitive and Computational Neuroscience, CTB, UPM, Madrid, Spain, **3** Dept. of Data Analysis, Faculty of Psychological and Educational Sciences, Ghent, Belgium, **4** Division of Computer Science, Universidad Pablo de Olavide, ES-41013 Seville, Spain, **5** Department of Informatics and Systems Engineering, University of La Laguna, Santa Cruz de Tenerife, Spain, **6** Unit of Clinical Neurophysiology, Teaching Hospital Ntra. Sra. de La Candelaria, Santa Cruz de Tenerife, Spain, **7** Department of Basic Medical Sciences, University of La Laguna, Santa Cruz de Tenerife, Spain

☯ These authors contributed equally to this work.
* eperdepa@ull.edu.es

Abstract

Functional connectivity (FC) characterizes brain activity from a multivariate set of N brain signals by means of an $N \times N$ matrix A, whose elements estimate the dependence within each possible pair of signals. Such matrix can be used as a feature vector for (un)supervised subject classification. Yet if N is large, A is highly dimensional. Little is known on the effect that different strategies to reduce its dimensionality may have on its classification ability. Here, we apply different machine learning algorithms to classify 33 children (age [6-14 years]) into two groups (healthy controls and Attention Deficit Hyperactivity Disorder patients) using EEG FC patterns obtained from two phase synchronisation indices. We found that the classification is highly successful (around 95%) if the whole matrix A is taken into account, and the relevant features are selected using machine learning methods. However, if FC algorithms are applied instead to transform A into a lower dimensionality matrix, the classification rate drops to less than 80%. We conclude that, for the purpose of pattern classification, the relevant features should be selected among the elements of A by using appropriate machine learning algorithms.

1 Introduction

Machine learning algorithms, and their approach to data mining ranging from pattern recognition to classification, provide relevant tools for the analysis of neuroimaging data (see [1–12]

for recent reviews and examples with different neuroimaging modalities and pathologies). Indeed, modern technologies such as magnetic resonance imaging (MRI), magnetoencephalography (MEG) and electroencephalography (EEG) generate an enormous amount of data per subject in a single recording session, which call for exactly these kind of algorithms to extract relevant information for applications such as, e.g., categorical discrimination of patients from matched healthy controls or prediction of individual (clinical and non clinical) variables. A salient feature of all these neuroimaging modalities, however, is that (specially in the case of MRI) the number of features, *p* (anatomical voxels for MRI) is huge (of the order of many thousands), whereas the number of subjects *n* is normally small (typically, two orders of magnitude smaller. See [13] for a recent review). This problem, which is a well-known issue in practical machine learning applications, is termed the *small-n-large-p* effect [14], which aggravates the *curse of dimensionality* associated to this data. Indeed, singling out the (possibly few) relevant features from the many thousands available has been compared to finding a needle in a haystack [7].

In the case of MRI, one way of tackling this issue consists in defining the so-called *regions of interest* (ROIs), an approach whereby the many voxels of the MRI are grouped to produce atlases, i.e., a coarser parcellation of the brain image. ROIs can be defined *ad hoc* or using some criterion such as cytoarchitectonics [15] (structure and organization of the neurons), as it is the case for the classical Broadmann areas of the cerebral cortex. In the case of MEG (and specially of the EEG), this problem is not so serious. In these two neuroimaging modalities, the number of recording sites (sensors for the MEG, or electrodes for the EEG) reduces to at most a few hundred, which, although still large and normally higher than the number of subjects, it is an order of magnitude lower than for the case of the MRI. Besides, and contrary to MRI, M/EEG present the advantage of a much higher temporal resolution (of the order of milliseconds), which allows characterizing diseases where one of the relevant features is the impairment of brain oscillatory activity at frequencies > 1 Hz. Thus, it may seem appealing to turn to these two modalities, where the curse of dimensionality is somehow controlled, for machine learning applications.

Recently, the study of brain activity from M/EEG has benefited from the development of new multivariate analysis techniques that characterizes the degree of functional (FC) and/or effective brain connectivity between two neurological time series (see [16, 17] for reviews). The application of these new techniques entails a paradigm shift, in which cognitive functions are no longer associated to specific brain areas, but to networks of interrelated, synchronously activated areas, networks that may vary dynamically to meet different cognitive demands [18, 19]. The interest of this approach has been confirmed by many studies, which have found that these brain networks are disrupted in many neurological diseases as compared to the healthy state [20–22]. Thus, it is not surprising that machine learning algorithms have been recently combined with M/EEG connectivity analysis to classify subjects as healthy controls or patients suffering from different diseases such as Alzheimer's [8], epilepsy [10, 11] and Attention Deficit Hyperactivity Disorder (ADHD) [2], and to identify EEG segments with the subjects that generate them [23].

However promising this combination may be, the problem with it is that the number of features is no longer bounded by that of sensors, but instead, for *N* recording sites (whether sensor or electrodes), one has $\mathcal{O}(N^2)$ features, which leads us back to the *small-n-large-p* pathway. Therefore, should we want to use machine learning algorithms for M/EEG connectivity patterns, it is almost compulsory to apply some type of algorithm, such as feature selection, which reduces the dimensionality of the problem. Yet such reduction can be carried out following different strategies. Indeed, there are two main options. One can, on the one hand, using some truly multivariate method to the connectivity patterns to reduce the dimensionality of the

feature vector. In this way, features are not actually selected, but mapped to a lower dimensional space using a suitable transformation. On the other hand, features can be selected by means of a machine learning algorithm, which favours those features most relevant for classification.

In this work we apply a well-known machine learning algorithm, the Bayesian Network Classifiers [24] (BNC) to classify 33 children into two different groups (healthy controls and ADHD) from their functional brain connectivity EEG patterns, obtained by using two indices of phase synchronisation (PS). ADHD is a well-known disorder, which has received a lot of attention recently in this framework [13, 25, 26]. Normally, the theta/beta power spectral ratio is used as the (already FDA supported) biomarker of reference to be used as adjunct to clinical assessment of such disease, although the latest literature [13] indicates that things may not be so clear-cut. Besides, little is known [2, 13, 25] about the possibility of using EEG-based FC methods for this purpose (and the best strategy thereof).

Therefore, we compare here the results obtained when applying to these subjects the two different approaches for dimensionality reduction mentioned above: one acting *a priori* on the connectivity patterns, whereby we go back to the "original" scenario with one feature per recording site (i.e., to $\mathcal{O}(N)$), and the other one a "traditional" machine learning feature selection algorithm whereby a subset of the $\mathcal{O}(N^2)$ features is selected based on their redundancy. Specifically, we chose the Fast Correlation Based Filter [27] (FCBF), a fast and efficient algorithm capable of capturing non-linear relationships between features, and the population-based Scatter Search (SS) algorithm [28], which uses a reference set composed of high-quality and dispersed solutions that evolves by combining them.

Finally, we also study the influence on the classification accuracy of different strategies to select the data segments.

We aim at finding out which combination of PS indices and strategy for dimensionality reduction of the feature vector is optimal for classification from FC patterns of scalp EEG data for this data set, in the hope that our results may be useful for other researchers applying the same approach to M/EEG data in different pathologies.

2 Methods

2.1 Subjects and EEG recording

The data set analysed here is a subset of a larger one described elsewhere [2], thus we only provide a brief account of its most important features. Two groups of subjects between 6 and 10 years old were selected for the study. The first one (patient group) consists of 19 boys suffering from ADHD of combined type, (mean age: 8 ± 0.3 y.), recruited from the Pediatric Service (Psychiatric branch) of the Hospital N.S. La Candelaria in Tenerife. Only subjects meeting ICD-10 criteria of Hyperkinetic Disorder [29] or DSM-V criteria of ADHD combined type [30] were included. The second one (control group) consists of 14 boys (mean age: 8.1 ± 0.48 y.) recruited among the children of hospital staff. Inclusion in any group was voluntary and written informed consent of the subject and his parents/guardians was obtained. The Ethical Committees of the University of La Laguna and of the University Hospital N.S. La Candelaria approved the study protocol, which was conducted in accordance with the Declaration of Helsinki.

EEG recordings lasting approximately one and a half hourwere carried out with the subjects at rest in a soundproof, temperature- and lighting-controlled, and magnetically and electrically shielded room in the clinical neurophysiology service of the hospital. The EEG (sampling rate, 256 Hz) was recorded with open (EO) and closed eyes (EC) between 12:00 and 14:00 using an analogical—digital Nihon Kohden Neurofax EEG-9200 with a channel cap according to the
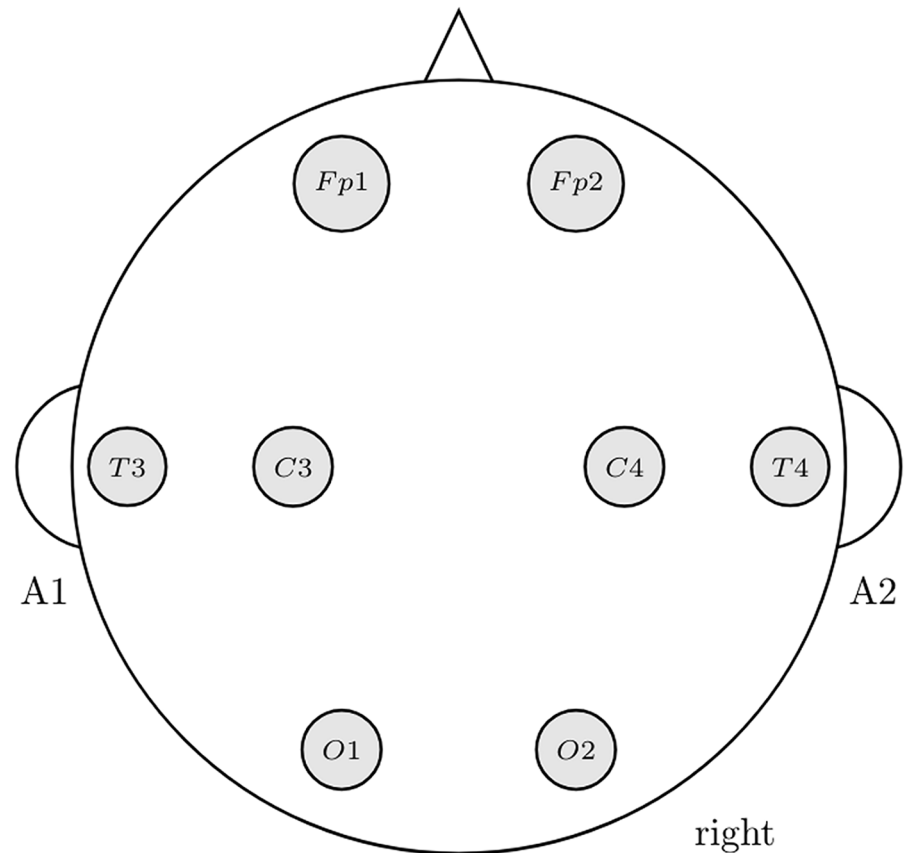
Fig 1. Electrode positions used in our experiments.

International 10/20 extended system, from the following downsampled set of eight channels (see Fig 1): Fp1, Fp2, C3, C4, T3, T4, O1 and O2. Each electrode was referenced to the contra-lateral ear lobe. Its impedance was monitored for each subject with the impedance map of the EEG equipment prior to data collection, and kept within a similar range (3 –5 kΩ). The data were filtered online using a high pass (frequency cut-off: 0.05 Hz), a low pass (frequency cut-off: 80 Hz) and a notch filter (50 Hz). Additionally, electro-oculograms and abdominal respiration movements were recorded for artefact detection.

2.2 Selection of the data segments

After discarding, by visual inspection, all the segments containing artefacts, the remaining data was divided into non-overlapping segments of 20s. (5, 120 samples), which were detrended and subsequently normalized to zero mean and unit variance. Then, we estimate the stationarity of each segment by calculating the average *ks* statistic of the Kwiatkowski—Phillips—Schmidt—Shin (KPSS) test for stationarity [31], as implemented in the GCCA toolbox [32]. Concretely, for each segment and subject we calculated:

$$\hat{ks^g} = \frac{1}{8}\sum_{i=1}^{8} ks_i^g \qquad (1)$$

where ks_i^g is the *ks* statistic of electrode *i* for segment *g*. The lower the value of (1), the lower the probability of segment *g* to be trend or mean non-stationary [31]. Therefore, we sorted the

values of (1) in ascending order ($k\hat{s}^{g_1} < k\hat{s}^{g_2} < ..$) and took the five segments $g_1, g_2, \ldots, g_5$ with the lowest values of the statistic, which are the five most stationary ones among all those available for each subject.

Finally, and prior to the estimation of the FC patterns (see below), the selected data segments were filtered using a Finite Impulsive Response (FIR) filter of zero phase distortion (filter order: 256) in the following five frequency bands: δ [$0.5 - 3.5 Hz$), θ [$3.5 - 8 Hz$), α [$8 - 13 Hz$), β [$13 - 30 Hz$) and γ [$30 - 48 Hz$).

2.3 Data analysis

2.3.1 Phase synchronisation analysis.   Phase synchronisation (PS) refers to a type of synchronized state in which the phases of two variables are locked, whereas their amplitudes are uncorrelated (see [33] for details, and, e.g., [16, 34] for a review of neuroscientific applications). The first step to study PS between two noisy real-valued signal consists of estimating the phases of each signal, which can be done in different ways [35]. We make use here of the approach based on the analytic signal $x_a(t)$, of a narrow band signal $x(t)$, which is constructed as follows:

$$x_a(t) = x(t) + jx_H(t) \tag{2}$$

where j is the imaginary unit ($j = \sqrt{-1}$) and $x_H(t)$ is the Hilbert transform of $x(t)$

$$x_H(t) = \frac{1}{\pi} P.V. \int \frac{x(\tau)}{t - \tau} d\tau \tag{3}$$

and P.V. stands for principal value. The phase of (2) is:

$$\theta_x(t) = \arctan \frac{x_H(t)}{x(t)} \tag{4}$$

The relative phase (restricted to the interval [$0, 2\pi$)) between electrodes i and l is defined as:

$$\varphi_{il}(t) = |\theta_{x_i}(t) - \theta_{x_l}(t)| \, mod \, 2\pi \tag{5}$$

The most usual way of assessing PS is the so-called *Phase Locking Value (PLV)*, defined as:

$$PLV_{il} = | < e^{j\varphi_{il}(\tau)} > | \tag{6}$$

where $< >$ indicates average, and $\| \|$ the norm of the resulting complex number. By definition, (6) ranges between 0 (no PS, or uniformly distributed φ_{il}) and 1 (complete PS or constant φ_{il}). It is closely related to the well-known coherency function, but taking into account only phase (rather than amplitude) information, and can be estimated very efficiently [36]. A well-known feature of this index when applied to scalp EEG is that it is unable to distinguish true connectivity, which takes place with non-zero time delay [37], from spurious FC between two electrodes recording the activity of a single deep neural source due to *volume conduction*), which is characterized by zero time delay.

Since the existence of a time delay in the interdependence gives rise to a relative phase centred around values other from 0 and π. Thus, a variant of (6) has been defined, which is robust against volume conductions effects by ignoring these two relative phase differences [38]:

$$PLI_{il} = | < sign(sin(\varphi_{il}(t)) > | \tag{7}$$

where $sign(x) = 1$ if $x > 0$, -1 otherwise. Clearly, (7) is 0 if the distribution of (5) is symmetric around 0 or π.

If we were only interested in the patterns of true connectivity with delay between pairs of N electrodes, (7) would be the appropriate choice (see, e.g., [39] for a recent example). But for the present application we think that there are good reasons to prefer the combined use of both indices, PLV_{il} and PLI_{il}. Firstly, it is well-known that indirect, yet neurologically meaningful connections, between two cortical networks via thalamic relay also take place with zero time lag, [40], which make them indistinguishable from volume conduction in this regards (see also [41]). Secondly, from the point of view of characteristic patterns, the fact that PLV is sensitive to the activity of deep brain sources is actually an advantage rather than a problem. Thus, we use both indices (as implemented in the recently released HERMES toolbox [42]) to characterize the patterns of brain dynamics of each subject, as explained in section 2.4.1.

2.3.2 Multivariate surrogate data test. It is well-known that the values of any of the PS indices described in Section 2.3.1, when applied to two finite-size, noisy experimental time series, may be affected by features of the data other than the existence of statistical relationships between them. In order words, one may have, e.g., that $PLV_{i,k} > 0$ even though $x_i(t)$ and $x_l(t)$ are actually independent from each other. To tackle this problem, it is advisable to estimate the significance of the PS indices before applying any classification algorithm. Here, we made use of the (bivariate) surrogate data method [43], whereby the original value of a FC index (say, PLV_{il} is compared to the distribution of N_s indices calculated from surrogates versions of x_i and x_l that preserve all their individual features (amplitude distribution, power spectrum..) but are independent by construction. Such *surrogate* signals can be generated in different ways [44, 45]. The simplest strategy in PS analysis consists of estimating the Fourier transform of the signals, add to the phase of each frequency a random quantity drawn from a uniform distribution between 0 and 2π and then transform them back to the time domain. In this way, any possible PS between the original signals is destroyed, but, as it turns out, this also destroys any coherent phase relationship present in each individual signal due to the nonlinearity of the system that generates it [44]. Since such nonlinearity cannot be ruled out in the case of EEG data (see, e.g., [46, 47]), more sophisticated algorithms are necessary. Thus, we chose the twin surrogate algorithm [45, 48, 49], which allows to test for phase synchronisation of complex systems in the case of passive experiments in which some of the signals involved may present nonlinear features. This algorithm works on the recurrence plot obtained from the signal, and is parametric, because it requires, for the proper reconstruction of the state space of the systems that generates the data, the embedding dimension m, which we estimated by using the false nearest neighbor method [50] and the delay time τ, which we took as the first minimum of the mutual information.

In this way, we generated $N_s = 99$ pairs of surrogate data $\{x_l, x_i^s\}$ ($s = 1, .., 99$), and estimated the distribution of PLV_{il} and PLI_{il} under the null hypothesis of no PS by calculating the corresponding PS index between each x_i^s and x_l. Finally, the original value of the index (say, PLV) was considered significant, at the p<0.01 level, if $PLV_{il} > PLV_{il}^s$ $\forall s$.

2.4 Classification

2.4.1 Feature vectors. The process of band pass filtering and PS assessment described before gives rise to FC matrices of the following form:

$$A_R^b = \begin{pmatrix} a_{R_{11}}^b & a_{R_{12}}^b & \cdots \\ a_{R_{21}}^b & a_{R_{22}}^b & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{8}$$

where R is either PLV or PLI, $b = \delta, \theta, \alpha, \beta, \gamma$ stands for each of the five frequency bands, and $1 = F3, 2 = C3,\ldots, 8 = O4$ are the electrodes as depicted in Fig 1.

Considering that both PS indices are symmetric (i.e., $a^b_{R_{il}} = a^b_{R_{li}} \forall i, l, R$), and that the diagonal values convey no information ($a^b_{R_{ii}} \equiv 1 \forall i$), the total number of features per band and index is $N_F = N(N-1)/2$, where $N (= 8$ in our case) is the number of electrodes analysed, i.e., $N_F = 28$. In other words, each feature vector per band b and index R is:

$$A^b_R = (a^b_{R_{12}}, \ldots, a^b_{R_{18}}, a^b_{R_{23}} \cdots,$$
$$\ldots, a^b_{R_{28}}, a^b_{R_{34}}, \ldots, a^b_{R_{68}}, a^b_{R_{78}}) \tag{9}$$

The whole procedure of feature vector construction is shown in Fig 2 for the case of the α band.

If we merge the twenty vectors such as (9) (one per band and condition for both *PLV* and *PLI*), we end up with a feature vector (recently termed as the *FCprofile* [51]) of 28x20 = 560 features:

$$\mathbf{R} = \{A^{O,\delta}_{PLV}, A^{O,\theta}_{PLV}, \ldots, A^{O,\gamma}_{PLV}, A^{O,\delta}_{PLI},$$
$$\ldots, A^{O,\gamma}_{PLI}, \ldots, A^{C,\delta}_{PLV}, A^{C,\theta}_{PLV}, \ldots, A^{C,\gamma}_{PLI}\} \tag{10}$$

where the superscripts O and C stand for open and closed eyes, respectively.

Note that, if one does not apply the surrogate data test, $a^b_{R_{il}} > 0 \forall i, l, b, R$ and both conditions. After applying it, however, for a given subject on has $a^b_{R_{il}} = 0$ for those values of the index R that do not pass the test (say, e.g., $a^\alpha_{PLV_{23}}$ for open eyes, both but possibly not for closed
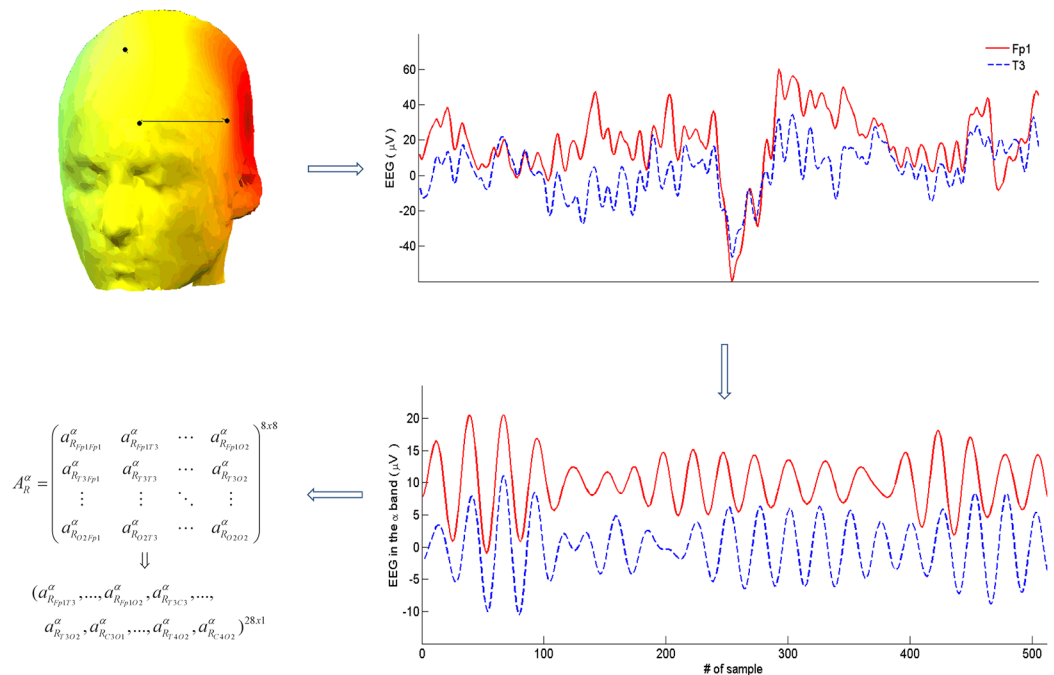
Fig 2. Schematic representation of the construction of the feature vector for each band and index. For each pair of channels as in (a), the raw data in (b) are filtered in the electrodes (Fp1 and T3 in this example), segments such as those in (b) are selected. Then, the signals are filtered in the corresponding frequency bands (e.g., α in (c)), and the 8×8 connectivity matrix A^α_R is obtained, which is finally converted to the 1×28 feature vector, after removing the diagonal elements and taking into account the symmetry of both PS indices (i.e., $a^b_{R_{ii}} = 1; a^b_{R_{il}} = a^b_{R_{li}} \forall i, l, b$ and R).

https://doi.org/10.1371/journal.pone.0201660.g002

eyes). Yet this method cannot be used to reduce the dimensionality of (10), because $a^\alpha_{PLV_{23}}$ may be indeed significant for another subject. Thus, in general, if we use for the purpose of FC pattern classification, R symmetric bivariate indices providing complementary information, calculated on N signals in b independent frequency bands, we end up with feature vectors such as (10), with $b \times R \times N(N-1)/2$ features. These are the feature vectors used by the machine learning classification algorithms described henceforth.

2.4.2 Bayesian Network Classifier (BNC). In machine learning, classification is the problem of learning a function that identifies the category to which a new observation belongs to. Formally, let $\mathcal{T}$ be a set composed by n instances described by pairs $(\mathbf{x}_i, y_i)$, where each $\mathbf{x}_i$ is a vector described by d quantitative features, and its corresponding y_i is a qualitative attribute that stands for the associated class to the vector. The classification problem consists of inducing a function $\mathcal{C} : \mathbf{X} \rightarrow \mathcal{Y}$ called classifier such that maps from a vector $\mathbf{X}$ to class labels $\mathcal{Y}$.

We use the BNC [24] due to its ability to explain the causal relationships among the features by using the joint probability distribution. These causal relationships allow to model correlation among the features as well as make predictions of the class label $\mathcal{Y}$. A BNC is a probabilistic graphical model that represents the features and their conditional dependencies as a directed acyclic graph (DAG). A node represents a feature and the edges represent conditional dependences between two features.

Building the classifier consists in learning the structure of the network that best fits the joint distribution of all features given the data, and the set of conditional probability tables (CPTs). Fig 3 shows a Bayesian network for binary data. As we can see, there is always an edge from the class variable $\mathcal{Y}$ to each feature X_i. The edge from X_2 to X_1 implies that the influence of X_1 on the assessment of the class variable also depends on the value of X_2. Structure learning is computationally very expensive and has been shown to be an NP-hard problem [52], even for approximate solutions [53]. Therefore, learning Bayesian networks normally requires the use of heuristics and approximate algorithms to find a local maximum in the structure space and a score function that evaluates how well a structure matches the data.

To learn the structure of the Bayesian networks, we used the following algorithms: K2 [54], Hill Climbing [55] and LAG Hill Climbing. K2 is a greedy search strategy that begins by
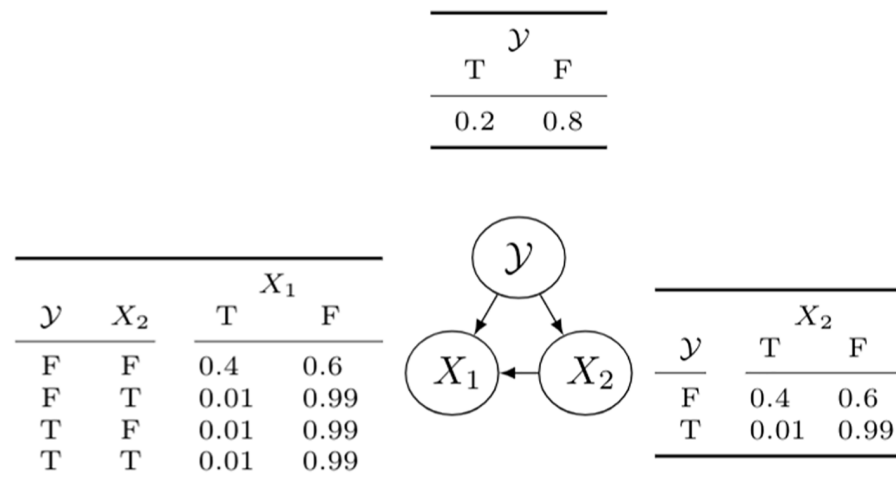
Fig 3. Bayesian network for binary features. Nodes represent features and edges conditional dependencies. The model specifies the conditional Probability Table (CPT) for each feature, which lists the probability that the child node takes on each of its different values for each combination of values of its parents.

assuming that a node has no parents. Then, it adds incrementally that parent, from a given ordering, whose addition increases most the score of the resulting structure. This strategy stops when there is no increase of the score when adding parents to the node. Hill Climbing algorithms begin with an initial network and at each iteration apply single edge operation (adding, deleting and reversing) until reaching a locally optimal network. Unlike K2, the search is not restricted by an ordering of variables. LAG Hill Climbing performs hill climbing with look ahead on a limited set of best scoring steps. With the purpose of quantifying the fitting of the obtained Bayesian networks, we use the Bayesian Dirichlet [56] (BD) scoring function.

2.4.3 Dimensionality reduction. In principle, each of the components of the feature vector 10 offers information on the FC patterns. Thus, we would be faced with the problem of classifying a set of k subjects using n features, where $k \ll n$. Yet, it is not difficult to foresee that there are cases where there exists a high degree of redundancy between some or many of the such components. For instance, if the connection between the brain networks recorded by electrodes i and l at a given frequency is direct (or non-existent) then PLV_{il} and PLI_{ij} provides essentially the same information. Thus, it is reasonable, to lessen the so-called "curse of dimensionality" to apply some kind of procedure to reduce the number of useful (non-redundant) features to be used for classification. This aim can be accomplished in two different ways. The first one consists of selecting a subset of the available features by using feature selection algorithm from the field of machine learning, which allows maintaining the classification accuracy while minimizing the number of necessary features. The second one, which is specific to multivariate PS analysis, entails the derivation, from each of the matrices 8, of a reduced set of indices that summarize the information of the PS pattern at each frequency band by applying truly multivariate PS methods such as those described, e. g., in [57, 58]. Henceforth, we detail how both approaches were carried out.

Feature selection via machine learning algorithms

As commented above, in classification tasks the aim of feature selection is to find the best feature subset, from the original set, with the smallest lost in classification accuracy. The goodness of a particular feature subset is evaluated using an objective function, $J(S)$, where S is a feature subset of size $|S|$.

In our experiments we use, as feature selection algorithm, the Fast Correlation Based Filter [27] (FCBF). FCBF is an efficient correlation-based method that performs a relevance and redundancy analysis for selecting a good subset of features. It consist in a backward search strategy that uses Symmetrical Uncertainty (SU) as objective function to calculate dependences of features. Since SU is an entropy based non-linear correlation, it is suitable for detecting non-linear dependencies between features.

By considering each feature as a random variable, the uncertainty about the values of a random variable X is measured by its entropy $H(X)$, which is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \tag{11}$$

Given another random variable Y, the conditional entropy $H(X|y)$ measures the uncertainty about the value of X given the value of Y and is defined as

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(xi|y_j)\log_2(P(x_i|y_j)) \tag{12}$$

where $P(y_j)$ is the prior probability of the value y_j of Y, and $P(x_i|y_j)$ is the posterior probability of a given value x_i of variable X given the value of Y. Information Gain [59] of a given variable

X with respect to variable Y (IG(Y;X)) measures the reduction in uncertainty about the value of X given the value of Y and is given by

$$IG(X|Y) = H(X) - H(X|Y) \qquad (13)$$

Therefore, IG can be used as correlation measure. For example, given the r.v. X, Y and Z, X is considered to be more correlated to Y than Z, if $IG(Y|X) > IG(Z|X)$. IG is a symmetrical measure; which is a desired property for a correlation measure. However it is biased in favor of r.v. with more values and such values have to be normalized to ensure the values have the same scale and so are comparable and have the same effect. To overcome the bias drawback we use the Symmetrical uncertainty (SU) measure; which modifies IG measure by normalizing with their corresponding entropy to compensate the bias

$$SU(X, Y) = 2 \left[\frac{IG(X|Y)}{H(X) + H(Y)} \right] \qquad (14)$$

SU restricts its values to the range [0, 1]. A value of 1 indicates that knowing the values of either feature completely predicts the values of the other; a value of 0 indicates that X and Y are independent. So SU can be used as a correlation measure between features.

Based on SU correlation measure, the authors define the approximate Markov blankets as follows.

Definition 1 (Approximate Markov blanket) *Given two features X_i and X_j ($i \neq j$) so that $SU(X_j, \mathcal{Y}) \geq SU(X_i, \mathcal{Y})$, then X_j forms an approximate Markov blanket for X_i iff $SU(X_i, X_j) \geq SU(X_i, \mathcal{Y})$.*

To guarantee that a redundant feature removed in a given step will still find a Markov blanket in any later phase when another redundant feature is removed, they also introduce the concept of predominant feature.

Definition 2 (Predominant feature) *Given a set of features $S \subseteq \mathcal{X}$, a feature X_i is a predominant feature of S if it does not have any approximate Markov blanket in S.*

As we can see in Fig 4, it starts by calculating $SU(X_i, \mathcal{Y})$ for each feature to estimate the relevance. A feature is considered irrelevant if its value is lower or equal to a given threshold δ. In order to detect a subset of predominant features, remaining features are ordered in descending $SU(X_i, \mathcal{Y})$ value. Then a backward search is performed in the ordered list S'_{list} to remove redundant features. The first feature from S'_{list} is a predominant feature since it has no approximate Markov blanket. Note that a predominant feature X_j can be used to filter out other features for which X_j forms an approximate Markov blanket. Therefore a feature X_i is removed from S'_{list} if X_j forms a Markov blanket for it. The process is repeated until no predominant features are found. In this work we set $\delta = 0$ since there is no rule about this parameter tuning and in the datasets under study only a small subset of features have a SU value different to 0.

The second method we used for feature selection is based on the Scatter Search (SS) metaheuristic proposed by García et al. [28]. SS is a population-based algorithm that makes use of a subset of high quality and dispersed solutions, which are combined to construct new solutions. The pseudocode of SS is summarized in Fig 5. The method generates an initial population of solutions in line 1, which is composed of solutions dispersed in the solution space. In line 2, a reference set of high quality and dispersed solutions is generated from the population. As in standard implementations of SS, the *SelectSubset* method in line 5 selects all subsets consisting of two solutions, which are then combined in line 6. The resulting solutions are then improved in line 7 obtaining new local optima. Finally, a static update of the reference set is carried out in line 9, in which a new reference set is obtained from the union of the original set and all the

Procedure *Fast Correlation Based Filter*
begin
1:    **for** $i = 1$ to d **do**
2:        calculate SU_{ic} for X_i;
3:        **if** $(SU_{ic} > \delta)$
4:            append X_i to S'_{list};
5:    **end**;
6:    order S'_{list} in descending SU_{ic} value;
7:    $X_j = \text{getFirstElement}(S'_{list})$;
8:    **do begin**;
9:        $X_i = \text{getNextElement}(S'_{list}, X_j)$;
10:    **if**$(X_i <> NULL)$
11:        **do begin**;
12:            **if**$(SU_{ij} \geq SU_{ic})$;
13:                remove X_i from S'_{list};
14:                $X_i = \text{getNextElement}(S'_{list}, X_j)$;
15:        **end until** $(X_i == NULL)$;
16:        $X_j = \text{getNextElement}(S'_{list}, F_j)$;
17: **end until** $(X_j == NULL)$;
18: $S_{best} = S'_{list}$;
end

Fig 4. Pseudocode of the Fast Correlation Based Filter algorithm (FCBF).

combined and improved solutions by quality and diversity. For more details, we refer the interested reader to the original paper.

The novelty introduced in this paper is that, in order to measure the quality of the subsets of features selected by the scatter search, we made use of the Correlation Feature Selection (CFS) measure [60] instead of a wrapper approach. CFS evaluates subsets of features taking into account the hypothesis that the good subsets include features highly correlated with the classification, but uncorrelated to each other. It evaluates the worth of a subset of attributes by

Procedure *Scatter Search*

begin

1: CreatePopulation (*Pop*, PopSize);

2: GenerateReferenceSet (*RefSet*, RefSetSize);

3: **repeat**

4:   **repeat**

5:       SelectSubset (*SubSet*, RefSetSize);

6:       CombineSolutions (*SubSet*, CurSol);

7:       ImproveSolutions (*CurSol*, ImpSol);

8:   **until** *StoppingCriterion1*;

9:   UpdateReferenceSet (*RefSet*);

10: **until** *StoppingCriterion2*;

end

Fig 5. Pseudocode of the Scatter Search algorithm (SS).

https://doi.org/10.1371/journal.pone.0201660.g005

considering the individual predictive ability of each feature along with the degree of redundancy between them.

The subsest evaluation function of CFS can be stated as follows:

$$M_S = \frac{k \overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}},$$ (15)

where M_S is the heuristic merit of a feature subset S containing k features, $\overline{r_{cf}}$ is the mean feature-class correlation ($f \in S$), and $\overline{r_{ff}}$ is the average feature-feature inter-correlation.

Dimensionality reduction using FC methods: *synchronisation cluster algorithm*

The need for dimensionality reduction in FC studies was already recognized even before the possibility of using them in Machine Learning applications. Indeed, apart from the practical issues associated to the multiple comparison problem (see, e.g., [61] and references therein), it was also demonstrated that weak pairwise correlations (i.e., low values of $a^b_{R_{ij}}$) may indicate, rather counter-intuitively, strongly correlated neural network states [62]. In the specific case of PS analysis, Allefeld and co-workers [57] developed a method termed *synchronisation Cluster Analysis* (SCA) whereby the N electrodes/sensors from a multivariate M/EEG recording can be considered, under very general conditions, as individual oscillators coupled to a common oscillatory rhythm. The degree of coupling of each oscillator to this global rhythm, ρ_i, as well as the overall strength of the joint synchronized behaviour of all the oscillators, r, can be inferred from the matrix 8 (see [57] for details). Thus, the FC pattern for each PS index R and frequency band b comes down to a $\mathcal{O}(N)$ vector rather than to a $\mathcal{O}(N^2)$ one.

Namely:

$$\rho_R^b = (\rho_{R_1}^b, \rho_{R_2}^b, .., \rho_{R_n}^b) \tag{16}$$

Contrary to the approach described above, with SCA we actually do not select a subset of the $\mathcal{O}(N^2)$ features, but rather reduce the number of features to $\mathcal{O}(N)$ by taking advantage of the characteristics of the dynamics of the brain synchronisation as described by the data. Note that this approach is also equivalent to another FC methods of dimensionality reduction, such as those based on describing 8 in terms of its N eigenvalues [58] or the more recent one based on hyperdimensional geometry [63]. For the 5 frequency bands considered, the two PS indices and the two conditions (open and closed eyes) we have, for each subject, a feature vector ρ of N(=8)×5×2×2 = 160 components. Finally, if we take the reduction approach to the limit, and use r instead of (16), then the feature vector comprises only 20 components, one for each possible band / index /condition combination.

3 Experiments and results

For each subject, we selected three different sets of feature vectors (R, R^t and R_S^t), with the aim of determining the influence of each processing steps on classification accuracy. Thus, R stands for the feature vectors obtained from 5 segments selected randomly out of all the available one. Then, R^t corresponds to the results from the 5 most stationary segments with the selection procedure described in section 2.2. Finally, R_S^t is the same that R^t but applying the surrogate data test. Let ρ and r and ρ^t and r^t be the datasets obtained by applying the SCA algorithm to R and R^t, respectively. As for R_S^t, there are instances in which the matrix (8) is sparse (i.e., there are many non-significant indices), which prevents the application of the SCA algorithm. Therefore, it was not possible to calculate either ρ_S^t or r_S^t.

In Table 1, we present the main characteristics of the datasets used in the experiment. The first column refers to the set of feature vector. The following two columns show the datasets obtained from the feature vectors depending on whether SCA was applied or not. Then, for each dataset, the number of features per band is presented and finally, in the last column, we can see the total number of features. Datasets are generated for PLI and PLV phase synchronisation methods. Note that for each band and phase synchronisation method, we include measures from two eyes positions (open and closed).

To evaluate and compare the predictive models learned from data, we used cross-validation; which is a popular method for estimating generalization error based on re-sampling and thus assesses model quality.

In cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. The basic form of cross-

Table 1. Characteristics of the different datasets used in this work for PLI and PLV phase synchronisation methods.

data	FC	dataset	#features/band	#features
R	sca	r	2	10
		ρ	16	80
	–	R	28	280
R^t	sca	r^t	2	10
		ρ^t	16	80
	–	R^t	28	280
R_S^t	–	R_S^t	28	280

validation is k-fold cross-validation, which splits the data into k equally sized subsets or folds. Then, k iterations of training and validation are performed such that each time a different fold is held-out for validation while remaining $k − 1$ folds are used for learning purpose. Finally, the validation results are averaged over the runs. In general, lower values of k produce more pessimistic estimates and higher values more optimistic ones. However, since the *true* generalization error is not usually known, it is not possible to determine whether a given result is an overestimate or underestimate. In spite of this, cross-validation is a suitable estimator for model comparison purposes.

Although cross-validation consumes a great deal of resources, for small sized data the leave-one-out cross-validation (LOOCV) is used since it is an almost unbiased estimator, albeit with high variance. LOOCV is a special case of k-fold cross-validation, where k equals the number of instances in the data.

All EEG data used in this work are bi-class (e.g., the subjects are either control or ADHD), so that we use sensitivity and specificity scores as performance measures. In our data positive examples refer to label ADHD while negative to control cases. Sensitivity, also called *true positive rate* or recall, measures the proportion of actual positives which are correctly identified as such. Higher values means that higher cases of ADHD are detected. Specificity is the proportion of actual negatives which are identified as such. Higher values correspond to lower probability that a control case be classified as ADHD case.

3.1 Baseline classification results

In this section, we analyse the predictive power of the different search strategies for Bayes network structure learning with the datasets under study. Table 2 presents the results. The phase synchronisation method applied is indicated in the first column. Then, the dataset id is shown. The following columns refer to sensitivity and specificity scores for K2, Hill Climbing (HC), and LAG Hill Climbing (LHC). Results with an accuracy higher than 0.7 are in bold.

With the PLI index, K2 achieves an accuracy higher than 0.70 on R dataset. With PLV index, results on R^t dataset achieves a very high accuracy with all search strategies. The other results are lower than 0.70.

Table 2. Sensitivity and specificity obtained with K2, HC, and LHC search strategies for Bayes network structure learning. Results with accuracy values higher than or equal to 0.70 are marked in bold.

classifier		K2		HC		LHC	
PS	**id**	**sens.**	**spec.**	**sens.**	**spec.**	**sens.**	**spec.**
PLI	r	0.421	1.000	0.421	1.000	0.421	1.000
	ρ	0.421	0.875	0.421	0.813	0.421	0.813
	R	**0.684**	**0.733**	0.684	0.600	0.684	0.600
	r^t	0.947	0.000	0.947	0.000	0.947	0.000
	ρ^t	0.684	0.467	0.684	0.467	0.684	0.467
	R^t	0.526	0.467	0.526	0.467	0.526	0.467
	R_S^t	1.000	0.000	1.000	0.000	1.000	0.000
PLV	r	1.000	0.000	1.000	0.000	1.000	0.000
	ρ	1.000	0.000	1.000	0.000	1.000	0.000
	R	0.579	0.667	*na*	*na*	0.526	0.667
	r^t	0.579	0.000	0.579	0.000	0.579	0.000
	ρ^t	0.421	0.733	0.421	0.733	0.421	0.733
	R^t	**0.895**	**0.933**	**0.895**	**0.867**	**0.947**	**0.933**
	R_S^t	0.474	0.667	0.474	0.667	0.474	0.667

Table 3. Sensitivity and specificity obtained with K2, HC, and LHC search strategies for Bayes network strcuture learning after preprocessing with FCBF and SS feature selection algorithms. Results with accuracy values higher than or equal to 0.70 are marked in bold.

| PS	id	FCBF						SS					
		K2		HC		LHC		K2		HC		LHC	
		sens.	spec.	sens.	spec.	sens.	spec.	sens.	spec.	sens.	spec.	sens.	spec.
PLI	r	0.421	1.000	0.421	1.000	0.421	1.000	0.421	1.000	0.421	1.000	0.421	1.000
	ρ	0.421	0.875	0.421	0.875	0.421	0.875	0.421	0.875	0.421	0.938	0.421	0.875
	R	**0.684**	**0.733**	0.684	0.600	0.684	0.600	**0.684**	**0.733**	0.684	0.600	0.684	0.600
	r^t	0.947	0.000	0.947	0.000	0.947	0.000	0.947	0.000	0.947	0.000	0.947	0.000
	ρ^t	0.684	0.467	0.684	0.467	0.684	0.467	0.684	0.467	0.684	0.467	0.684	0.467
	R^t	0.526	0.467	0.526	0.467	*na*	*na*	0.526	0.467	0.526	0.467	*na*	*na*
	R^t_s	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
PLV	r	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
	ρ	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
	R	0.526	0.533	0.579	0.600	0.632	0.600	0.632	0.533	0.579	0.533	0.684	0.600
	r^t	0.579	0.000	0.579	0.600	0.579	0.000	0.579	0.000	0.579	0.000	0.579	0.000
	ρ^t	0.421	0.733	0.421	0.733	0.421	0.733	0.421	0.733	0.421	0.733	0.421	0.733
	R^t	**0.947**	**0.867**	**0.947**	**0.867**	**1.000**	**1.000**	**0.947**	**0.933**	**0.947**	**0.933**	**1.000**	**0.933**
	R^t_s	0.474	0.667	0.474	0.667	0.474	0.667	0.474	0.667	0.474	0.667	0.474	0.667

3.2 Feature selection analysis

Regarding the effect of the feature selection algorithms FCBF and SS on the classification performance, results are shown in Table 3. Those obtained by applying the FCBF are shown in columns 3–8 while those achieved with SS in columns 4–9. As in the baseline scenario, K2 achieves the same accuracy on *R* with PLI index, and all search strategies achieve the best performance scores on R^t using PLV index. With FCBF, the model found by K2 achieves the same accuracy by increasing the sensitivity and decreasing the specificity. The model found by HC improves the accuracy by increasing the sensitivity. Finally, the predictive power found by LHC achieves an accuracy of 100%. This results must be taken with certain caution since they may suggest some kind of overfitting. With SS, the model of K2 improves increases the sensitivity while remaining the specificity value. HC improves in both measures and LHC increases sensitivity reaching a value of 1.

3.3 Band relationship analysis

In this section we analyse the subsets of features selected by FCBF and SS on R^t with PLV. Fig 6 shows the features (connections from now on) selected according to the electrode positions used in our experiments. The connections selected by FCBF are shown in Fig 6A, while those selected by SS are in Fig 6B. The width of the connection is larger for those with higher correlation values with the class label. We used SU as a measure of feature correlation. Superscript *c* stands for connections measured with closed eyes while no superscript refers to open eyes. Along this section we will write the connection between two electrodes E1 and E2 in a given band as $(E1 - E2)_{band}$ for opened eyes and $(E1 - E2)^c_{band}$ for closed ones.

When applying FCBF, 12 out of the 280 features have non-zeron SU correlation values. Then, the strategy selected 10 of them as predominant features. The selected connections are shown in Fig 6A. The most correlated connections correspond to $(T3 - Fp2)_\beta$ with $SU = 0.571$ and $(O2 - C4)^c_\gamma$ with $SU = 0.536$. Most values are in the range of $[0.303 - 0.380]$ and only two
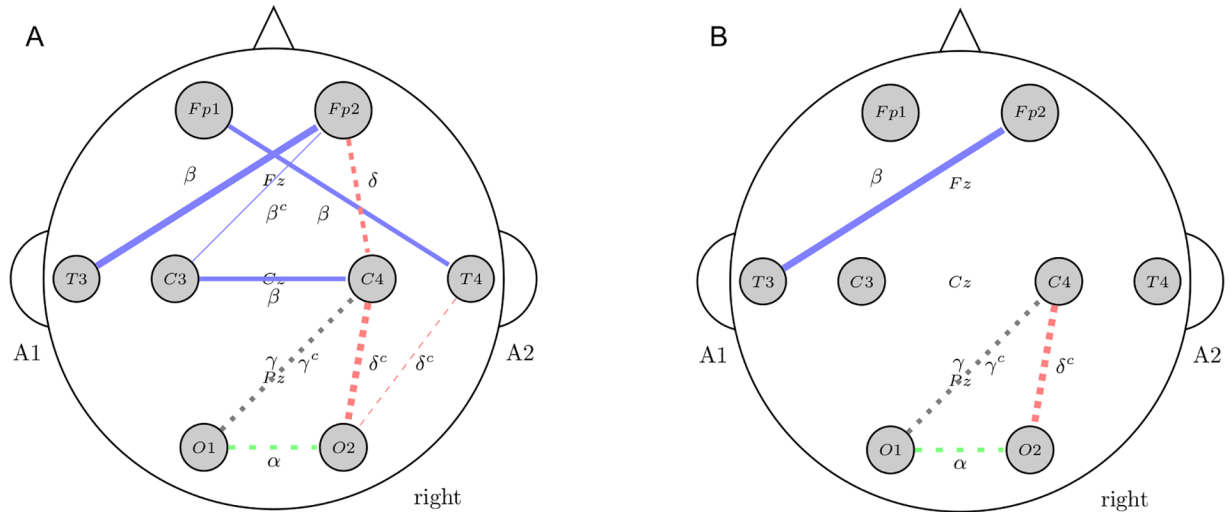
Fig 6. Connections (features) selected by (A) the FCBF and (B) SS feature selection algorithms on R^t dataset with PLV index. The type of line indicates the band (δ: dashed; β: solid; α: loosely dashed; γ: dotted), whereas its width is proportional to the correlation of the corresponding connections (see text for details). The superscript ^{c} on the letter for each band indicates the EC condition, whereas connections without superscript correspond to the EO condition.

of them have values lower than 0.3. The redundant connections found by this strategy are: $(Fp1 - O2)_\beta^c$ and $(C3 - T3)_\gamma^c$. The earlier one with $(O1 - O2)_\alpha$ and the later one with $(O1 - C4)_\gamma^c$.

As we can see in Fig 6B, SS found 5 features that were identified as predominant features by FCBF. It selects the two most correlated connections $(T3 - Fp2)_\beta$ and $(O2 - C4)_\gamma^c$, as well as the features $(O1 - C4)_\gamma$, $(O1 - C4)_\gamma^c$ and $(O2 - C4)_\delta^c$.

Now we will analyse the BN classifier models generated with the connections selected by FCBF and SS. Fig 7 shows the models obtained using the search methods K2, HC and LHC. As it was explained in this work, in the Bayesian model, edges represent conditional dependencies between the connections. Dashed lines stands for correlations between connections. Due to lack of space, a connection between two electrodes $E1$ and $E2$ in a given band is represented in the figure as $\binom{E1}{E2}_{band}$ for opened eyes cases and $\binom{E1}{E2}_{band}^c$ for closed ones. We can interpret the generated model as follows.

In Fig 7A we can see the model obtained with K2 with the features selected by FCBF. This model achieves an accuracy of 91.18%. Values of Connection $(O1 - C4)_\gamma$ depend on values of $(O1 - O2)_\alpha$. We can also see that connection $(C3 - Fp2)_\beta^c$ receives influence from $(T3 - Fp2)_\beta$ and $(C3 - C4)_\beta$ and $(T4 - O2)_\delta^c$ from $(C4 - O2)_\delta^c$ and $(C3 - C4)_\beta$. It is worth noting that $(C3 - C4)_\beta$ influences two different connections. Fig 7B shows the model with HC, which is quite similar to the previous one and achieves the same accuracy. In contrast to previous model, the dependency between connections $(O1 - C4)_\gamma$ and $(O1 - O2)_\alpha$ is inverted. Another difference is that $(C3 - C4)_\beta$ receives influences from $(T3 - Fp2)_\beta$ and $(C3 - Fp2)_\beta^c$. Finally, the model with LHC is presented in Fig 7C. The accuracy is 100% but the complexity of the model has increased considerably and, so, its interpretability. The new dependencies with respect to the previous models are that $(C3 - C4)_\beta$ and $(T4 - Fp1)_\beta$ are influenced by other four connections each. The influence of $(C3 - C4)_\beta$ comes from $(C3 - Fp2)_\beta^c$, $(T3 - Fp2)_\beta$, $(Fp2 - C4)_\delta$ and
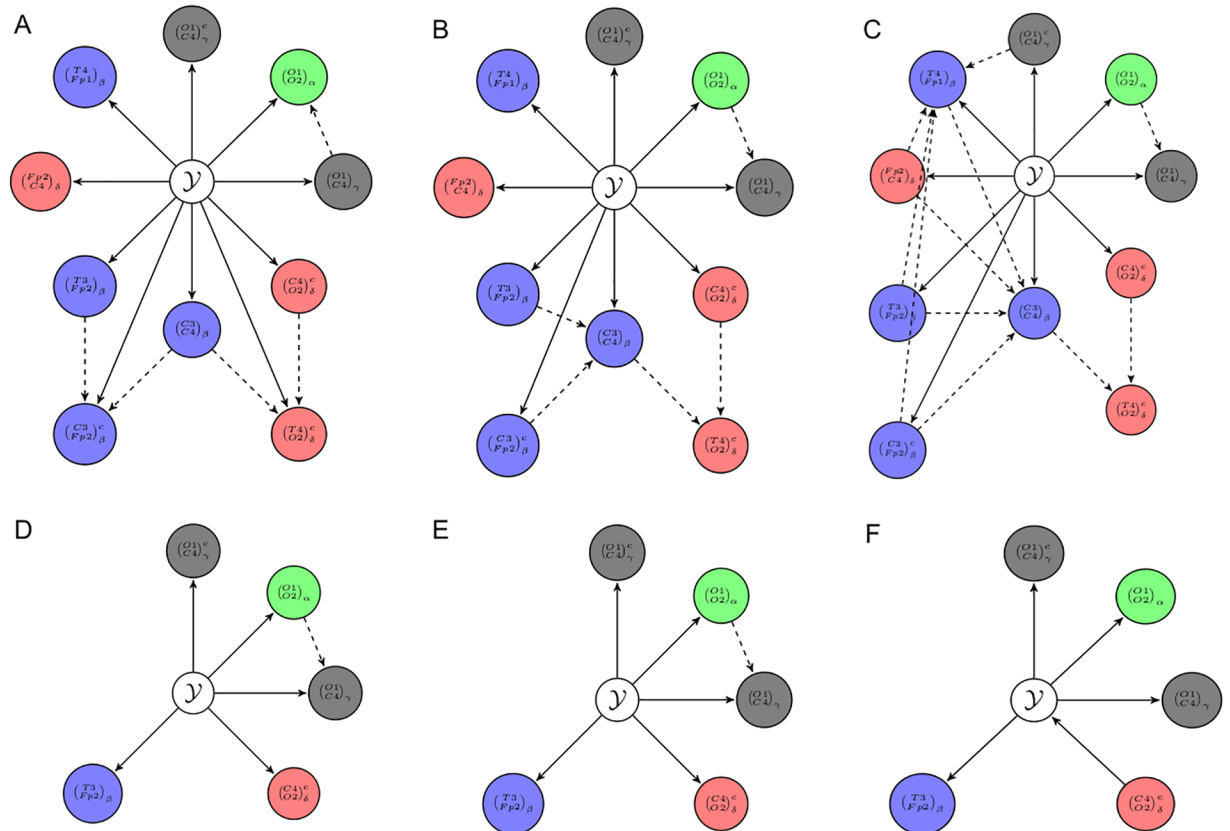
Fig 7. BNC models generated with the connections found by FCBF and SS. Dashed lines represent dependencies between such connections. (A) BNC model generates using K2 strategy with FCBF, (B) HC strategy and FCBF, (C) LHC algorithm and FCBF, (D) K2 and SS, (E) HC and SS, and (F) LHC and SS.

$(T4 - Fp1)_\beta$. Finally, $(T4 - Fp1)_\beta$ is influenced by $(O1 - C4)_\gamma^c$, $(Fp2 - C4)_\delta$, $(T3 - Fp2)_\beta$ and $(C3 - Fp2)_\beta^c$.

Note that in all models generated with connections selected by FCBF, we found that $(T4 - O2)_\delta^c$ depends on connections $(C3 - C4)_\beta$ and $(C4 - O2)_\delta^c$. The dependence between $(O1 - C4)_\gamma$ and $(O1 - O2)_\alpha$ is also present in all models although the direction of the arc is different in the first model.

The models obtained with SS are much simpler than those obtained with FCBF, since SS only selected five connections. As we can see in Fig 7D and 7E, K2 and HC algorithms learned the same BN model and it achieves an accuracy of 94.12%. Furthermore, this model shows a single statistical dependence between $(O1 - O2)_\alpha$ and $(O1 - C4)_\gamma$. Finally, in Fig 7F, we can see that the model obtained with LHC is slightly different to those obtained previously. It reaches the same accuracy (94.12%), but it presents no dependencies and connection $(C4 - O2)_\delta^c$ is the parent node of $\mathcal{Y}$. Therefore, if $\mathcal{Y}$ is known, $(C4 - O2)_\delta^c$ and the other four connections are conditionally independent.

Finally, it is noteworthy that the dependence between $(O1 - C4)_\gamma$ and $(O1 - O2)_\alpha$ is also presented in two of the three models generated with the connections selected by SS. This dependence is also found in the models generated previously with FCBF. Thus, we think that this is a robust result.

4 Discussion

We have analysed here a topic of great current interest [5, 9], namely the applicability of machine learning algorithms for subject classification from brain connectivity patterns. Concretely, we aimed at elucidating, using data from multichannel human EEG recordings, which is the best strategy to deal with the *curse of dimensionality* inherent to this research approach [14]. For this purpose, we compared different machine learning approaches of feature selection, which pinpoint the optimal subset of features out of all the available ones, with a method (SCA) based on modelling PS in brain dynamics to transform the original features in a reduced set of *new* variables. The whole procedure has been tested in a problem common in this framework, whereby brain connectivity is characterized using bivariate PS indexes between every two electrodes in different frequency bands [16, 19], and this feature vectors is then used to classify subjects in two groups [2, 7, 8, 10]. To the best of our knowledge, this is the first work where such a comparative study has been carried out.

Regarding the results of the classifier, we found that the combination of using the most stationary segments and the PLV yielded a high quality Bayesian network model. Additionally, the original FC features contain more information about the class than the transformed variables r and ρ. Such features correspond to the most informative connections for classification purposes in the brain connectivity pattern, since irrelevant and redundant ones are removed. Besides, as summarized in Fig 6, the application of FCBF/SS improves the interpretability of the classification model. In fact, Fig 6A and 6B present a very specific frequency/topology pattern of bands and electrodes whose FC, as assessed by PLV, is impaired in the ADHD groups as compared to the healthy one. Thus, low frequency activity in δ band is modified in the right hemisphere during CE condition, whereas higher frequency α, β and γ band FC changes mainly in the OE condition for interhemispheric connections. This is consistent with what is known about EEG activity in CE/OE conditions, where low frequency activity is enhanced in the former one, and also with the EEG changes associated to ADHD (see, e.g., [2, 25] and references therein). Note, however, that, as commented before, PLV and PLI measured different things [38], which justifies the use of both of them in the feature vectors. Yet, the best overall performance of any of the algorithms is obtained when only PLV features are considered. PLV is known to be more sensible to volume conduction effects, whereby the activity of a single neural source in the cerebral cortex or beneath is picked up by various electrodes, resulting in EEGs that are correlated. Quite interestingly, and also in line with recent results ([64, 65]), apparently this very reason turns this index into a richer source of information about the characteristic neuroimaging pattern of a given group, and correlates better with the underlying anatomical connectivity [41]. In other words, if one is not interested in the origin of the distinctiveness of the patterns but only wants to generate the most different patterns from two groups, then an index such as PLV, sensitive to changes both in the activity of deep sources and in the interdependence between them, may be more suitable for this purpose than the more robust PLI, which mainly detects the latter changes. Finally, given the inherent multiband nature of EEG changes in ADHD, it may be also interesting to use indices of cross-frequency coupling, which assess the interdependence between different frequencies (see, e.g., [66, 67]) to construct the FC vectors. Another interesting topic for further research would be the dynamic of FC patterns [26]. In both cases, the additional information would come at the price of further increasing the dimensionality, so that the present approach would be even more relevant there.

Another interesting issue that we have investigated is the effect of the segment selection procedure and the estimation of the statistical significance of the FC indices on classification accuracy. The main conclusion in this regard was that the optimal combination consists in

selecting the most stationary segments among those available ones (because randomly taking the segments always decreased the accuracy)and at the same time use the values of the indices as such, without any thresholding from the multivariate surrogate data. The result concerning the need for a careful selection of the segments according to their stationarity is hardly surprising. In fact, stationarity is known to be one of the prerequisites to estimate many interdependence indices such as correlation, coherence, mutual information or those based in the concept of generalized synchronisation [16], and the quantitative assessment of the degree of stationarity of M/EEG data segments in functional connectivity applications is receiving increasing attention [68]. Thus, even though stationarity is not a pre-requisite for the application of the Hilbert transform, it is anyhow quite logical that selecting stationary EEG segments that records a single brain state instead of non-stationary ones recording a mixture of them are the best candidates for pattern recognition applications. However, the poorer performance of the surrogate-corrected feature vectors as compared to the "raw" ones is somewhat surprising. Apparently, in the case analysed, the non-significant connectivity indices, whose values are not due to the statistical dependence between the time series but to some feature of the individual data (see, e.g.,[16, 43, 49]), do contain information that is relevant for the classification, so that setting all of them to zero produces more harm than good. Here, again, the message seems to be that the task of dimensionality reduction should be left to machine learning algorithms.

Admittedly, we cannot guarantee that these results will held for other sets of subject / neuroimaging modalities. It may be, for instance, that, contrary to what we have found here, there are instances where the original FC features, even after conveniently selected, do not outperform FC-based methods of dimensionality reduction such as SCA. Or that the use of surrogate data may be useful when the number of electrodes is high. Besides, it may be that BNC could no be always the best classifier. Different algorithms such as SVM [8], linear discriminant analysis [2] or random forest classifiers [10] have proven useful in similar applications. Note, however, that two of these works [8, 10] used previously transformed variables where data reduction is carried out by means of graph theoretic measures, and none of them perform a comparison of different classification algorithms or strategies for dimensionality reduction.

4.1 Limitations of the results

A very recent metanalysis of neuroimaging biomarkers in ADHD [13] has warned about the very high accuracies obtained in the literature in these type of studies. There, the small sample sizes and the circularity of the analysis, in which no cross-validation was used in many cases, were pointed out as the main causes of this inflated results. Although we did use cross-validation in the present study, it is clear that the size of our sample is small. Thus, rather that emphasising the absolute values of the accuracies obtained, we stress that they are the changes in this index (i.e., its relative values) after applying different approaches to select the segments and reduce the dimensionality of the feature vector, which represent most interesting outcome of our paper. Furthermore, by sharing all our data and making the code for connectivity analysis publicly available [42, 69], in line with recent efforts from our own research [12], we hope that other labs can apply the proposed classification model as build from our EEGs to their own data. This would be the best check for the validity of the proposed approach by estimating the accuracy of the model in external test sets, or alternatively would allow refining the model by enlarging the sample size. Although issues regarding the different pipelines may still be present, the detailed account we give on the preprocessing steps and automatic segment selection goes online with recent recommendations to improve reproducibility in neuroimaging research [70].

4.2 Conclusions

That said, our results suggest that the combination of a careful selection of the data segments, a suitable feature selection method and a machine learning algorithm such as BNC, able to cope with high dimensional data, can turn the curse of dimensionality into a blessing, as the availability of many features allows selecting an optimal subset of meaningful, information-rich variables that can accurately classify subjects from their brain connectivity patterns even from scalp EEG data.

In conclusion, the present outcomes indicate that the use of machine learning algorithms on EEG patterns of FC represents a powerful approach to investigate the underlying mechanisms contributing to changes, as regard to controls, in FC among different scalp electrodes, while allowing at the same time the use of this information for subject classification. They also suggest that this approach may not only be relevant for clinical applications (as it is the case for the theta/beta ratio in ADHD [13]), but also useful to provide insight into the neural correlates of the pathology under investigation.

Acknowledgments

Author Contributions

Conceptualization: Ernesto Pereda, Miguel García-Torres.

Data curation: Ernesto Pereda, Miguel García-Torres, Soledad Mañas, Leopoldo Méndez.

Formal analysis: Ernesto Pereda, Miguel García-Torres, Belén Melián-Batista.

Funding acquisition: Ernesto Pereda, Julián J. González.

Investigation: Miguel García-Torres, Belén Melián-Batista.

Methodology: Ernesto Pereda, Belén Melián-Batista, Julián J. González.

Resources: Julián J. González.

Software: Ernesto Pereda.

Supervision: Ernesto Pereda, Julián J. González.

Validation: Belén Melián-Batista, Soledad Mañas.

Writing – original draft: Ernesto Pereda, Miguel García-Torres, Julián J. González.

References

1. Bjornsdotter M. Machine Learning for Functional Brain Mapping. In: Application of Machine Learning. InTech; 2010. p. 147–170.

2. González JJ, Méndez LD, Mañas S, Duque MR, Pereda E, De Vera L. Performance analysis of univariate and multivariate EEG measurements in the diagnosis of ADHD. Clinical Neurophysiology. 2013; 6(124):1139–50.

3. Langs G, Rish I, Grosse-Wentrup M, Murphy B, editors. Machine Learning and Interpretation in Neuro-imaging—International Workshop, MLINI 2011, Held. Springer; 2012.

4. Richiardi J, Eryilmaz H, Schwartz S, Vuilleumier P, Van De Ville D. Decoding brain states from fMRI connectivity graphs. NeuroImage. 2011; 56(2):616–26. https://doi.org/10.1016/j.neuroimage.2010.05. 081 PMID: 20541019

5. Richiardi J, Achard S, Bunke H, Van De Ville D. Machine Learning with Brain Graphs: Predictive Modeling Approaches for Functional Imaging in Systems Neuroscience. IEEE Signal Processing Magazine. 2013; 30(3):58–70. https://doi.org/10.1109/MSP.2012.2233865

6. Shirer WR, Ryali S, Rykhlevskaia E, Menon V, Greicius MD. Decoding subject-driven cognitive states with whole-brain connectivity patterns. Cerebral Cortex. 2012; 22(1):158–65. https://doi.org/10.1093/cercor/bhr099 PMID: 21616982

7. Atluri G, Padmanabhan K, Fang G, Steinbach M, Petrella JR, Lim K, et al. Complex biomarker discovery in neuroimaging data: Finding a needle in a haystack. NeuroImage: Clinical. 2013; 3:123–31. https://doi.org/10.1016/j.nicl.2013.07.004

8. Zanin M, Sousa P, Papo D, Bajo R, García-Prieto J, del Pozo F, et al. Optimizing functional network representation of multivariate time series. Scientific Reports. 2012; 2:630. https://doi.org/10.1038/srep00630 PMID: 22953051

9. Zanin M, Papo D, Sousa PAA, Menasalvas E, Nicchi A, Kubik E, et al. Combining complex networks and data mining: why and how. Physics Reports. 2016 may; 635:58. https://doi.org/10.1016/j.physrep.2016.04.005

10. van Diessen E, Otte WM, Braun KPJ, Stam CJ, Jansen FE. Improved Diagnosis in Children with Partial Epilepsy Using a Multivariable Prediction Model Based on EEG Network Characteristics. PLoS ONE. 2013; 8(4):e59764. https://doi.org/10.1371/journal.pone.0059764 PMID: 23565166

11. Soriano MC, Niso G, Clements J, Ortín S, Carrasco S, Gudín M, et al. Automated Detection of Epileptic Biomarkers in Resting-State Interictal MEG Data. Frontiers in Neuroinformatics. 2017; 11. https://doi.org/10.3389/fninf.2017.00043 PMID: 28713260

12. Dimitriadis SI, López ME, Bruña R, Cuesta P, Marcos A, Maestú F, et al. How to Build a Functional Connectomic Biomarker for Mild Cognitive Impairment From Source Reconstructed MEG Resting-State Activity: The Combination of ROI Representation and Connectivity Estimator Matters. Frontiers in Neuroscience. 2018 jun; 12(JUN). https://doi.org/10.3389/fnins.2018.00306 PMID: 29910704

13. Pulini A, Kerr WT, Loo SK, Lenartowicz A. Classification accuracy of neuroimaging biomarkers in Attention Deficit Hyperactivity Disorder: Effects of sample size and circular analysis. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. 2018 jun;.

14. Mwangi B, Tian TS, Soares JC. A Review of Feature Reduction Techniques in Neuroimaging. Neuroinformatics. 2013, Epub ahead of print;.

15. Caspers S, Eickhoff SB, Zilles K, Amunts K. Microstructural grey matter parcellation and its relevance for connectome analyses. NeuroImage. 2013; 80:18–26. https://doi.org/10.1016/j.neuroimage.2013.04.003 PMID: 23571419

16. Pereda E, Quiroga RQ, Bhattacharya J. Nonlinear multivariate analysis of neurophysiological signals. Progress in Neurobiology. 2005; 77(1-2):1–37. https://doi.org/10.1016/j.pneurobio.2005.10.003 PMID: 16289760

17. Friston KJ. Functional and Effective Connectivity: A Review. Brain Connectivity. 2011; 1(1):13–6. https://doi.org/10.1089/brain.2011.0008 PMID: 22432952

18. Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience. 2009; 10(3):186–98. https://doi.org/10.1038/nrn2575 PMID: 19190637

19. Stam CJ, van Straaten ECW. The organization of physiological brain networks. Clinical Neurophysiology. 2012; 123(6):1067–87. https://doi.org/10.1016/j.clinph.2012.01.011 PMID: 22356937

20. Heuvel MP, Fornito A. Brain Networks in Schizophrenia. Neuropsychology Review. 2014;. https://doi.org/10.1007/s11065-014-9248-7 PMID: 24500505

21. Zhou J, Seeley WW. Network dysfunction in Alzheimer's disease and frontotemporal dementia: implications for psychiatry. Biological Psychiatry. 2014;. https://doi.org/10.1016/j.biopsych.2014.01.020

22. Maximo JO, Cadena EJ, Kana RK. The Implications of Brain Connectivity in the Neuropsychology of Autism. Neuropsychology Review. 2014, in press;. https://doi.org/10.1007/s11065-014-9250-0 PMID: 24496901

23. La Rocca D, Campisi P, Vegso B, Cserti P, Kozmann G, Babiloni F, et al. Human brain distinctiveness based on EEG spectral coherence connectivity; 2014.

24. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann; 1988.

25. Alba G, Pereda E, Mañas S, Méndez LD, González A, González JJ. Electroencephalography signatures of attention-deficit/hyperactivity disorder: clinical utility. Neuropsychiatric disease and treatment. 2015; 11:2755–69. https://doi.org/10.2147/NDT.S51783 PMID: 26543369

26. Alba E Guzmán and, Mañas S, Méndez LD, Duque MR, González A, González JJ. The variability of EEG functional connectivity of young ADHD subjects in different resting states. Clinical Neurophysiology. 2016 feb; 127(2):1321–30. https://doi.org/10.1016/j.clinph.2015.09.134 PMID: 26586514

27. Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research. 2004; 5:1205–24.

28. García-López F, García-Torres M, Melián-Batista B, Moreno-Pérez JA, Moreno-Vega JM. Solving feature subset selection problem by a Parallel Scatter Search. European Journal of Operational Research. 2006; 169(2):477–489. Cited By 69. https://doi.org/10.1016/j.ejor.2004.08.010

29. World Health Organization. WHO | ICD-10 classification of mental and behavioural disorders; 2002. Available from: http://www.who.int/substance_abuse/terminology/icd_10/en/.

30. López-Ibor JJ. DSM-IV-TR: manual diagnóstico y estadístico de los trastornos mentales. Barcelona: Masson, S.A.; 2002.

31. Kipiński L, König R, Sielużycki C, Kordecki W. Application of modern tests for stationarity to single-trial MEG data: transferring powerful statistical tools from econometrics to neuroscience. Biological Cybernetics. 2011; 105(3-4):183–95. https://doi.org/10.1007/s00422-011-0456-4 PMID: 22095173

32. Seth AK. A MATLAB toolbox for Granger causal connectivity analysis. Journal of Neuroscience Methods. 2010; 186(2):262–73. https://doi.org/10.1016/j.jneumeth.2009.11.020 PMID: 19961876

33. Rosenblum MG, Pikovsky AS, Kurths J. Phase synchronization of chaotic oscillators. Physical Review Letters. 1996; 76(11):1804–7. https://doi.org/10.1103/PhysRevLett.76.1804 PMID: 10060525

34. Bastos AM, Schoffelen JM. A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. Frontiers in Systems Neuroscience. 2016; 9. https://doi.org/10.3389/fnsys.2015.00175 PMID: 26778976

35. Bruns A. Fourier-, Hilbert- and wavelet-based signal analysis: are they really different approaches? Journal of Neuroscience Methods. 2004; 137:321–2. https://doi.org/10.1016/j.jneumeth.2004.03.002 PMID: 15262077

36. Bruña R, Maestú F, Pereda E. Phase Locking Value revisited: teaching new tricks to an old dog. arXiv. 2017;.

37. Nolte G, Bai O, Wheaton L, Mari Z, Vorbach S, Hallet M. Identifying true brain interaction from EEG data using the imaginary part of coherency. Clinical Neurophysiology. 2004; 115(10):2292–307. https://doi.org/10.1016/j.clinph.2004.04.029 PMID: 15351371

38. Stam CJ, Nolte G, Daffertshofer A. Phase lag index: Assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources. Human Brain Mapping. 2007; 28(11):1178–93. https://doi.org/10.1002/hbm.20346 PMID: 17266107

39. Hillebrand A, Barnes GR, Bosboom JL, Berendse HW, Stam CJ. Frequency-dependent functional connectivity within resting-state networks: An atlas-based MEG beamformer solution. NeuroImage. 2012; 59(4):3909–21. https://doi.org/10.1016/j.neuroimage.2011.11.005 PMID: 22122866

40. Gollo LL, Mirasso C, Villa AEP. Dynamic control for synchronization of separated cortical areas through thalamic relay. NeuroImage. 2010; 52(3):947–55. https://doi.org/10.1016/j.neuroimage.2009.11.058 PMID: 19958835

41. Finger H, Bönstrup M, Cheng B, Messé A, Hilgetag C, Thomalla G, et al. Modeling of Large-Scale Functional Brain Networks Based on Structural Connectivity from DTI: Comparison with EEG Derived Phase Coupling Networks and Evaluation of Alternative Methods along the Modeling Path. PLOS Computational Biology. 2016 aug; 12(8):e1005025. https://doi.org/10.1371/journal.pcbi.1005025 PMID: 27504629

42. Niso G, Bruña R, Pereda E, Gutiérrez R, Bajo R, Maestú F, et al. HERMES: Towards an Integrated Toolbox to Characterize Functional and Effective Brain Connectivity. Neuroinformatics. 2013; 115(11):405–34. Available from: http://hermes.ctb.upm.es/. https://doi.org/10.1007/s12021-013-9186-1

43. Andrzejak RG, Kraskov A, Stogbauer H, Mormann F, Kreuz T. Bivariate surrogate techniques: Necessity, strengths, and caveats. Physical Review E. 2003; 68(6):66202. https://doi.org/10.1103/PhysRevE.68.066202

44. Schelter B, Winterhalder M, Timmer J, Peifer M. Testing for phase synchronization. Physics Letters A. 2007; 366(4-5):382–390. https://doi.org/10.1016/j.physleta.2007.01.085

45. Thiel M, Romano MC, Kurths J, Rolfs M, Kliegl R. Twin surrogates to test for complex synchronization. Europhysics Letters. 2007; 75(4):535–41. https://doi.org/10.1209/epl/i2006-10147-0

46. Pereda E, Rial R, Gamundi A, Gonzalez J. Assessment of changing interdependencies between human electroencephalograms using nonlinear methods. Physica D. 2001; 148(1-2):147–58. https://doi.org/10.1016/S0167-2789(00)00190-1

47. de la Cruz DM, Mañas S, Pereda E, Garrido JM, Lopez S, De Vera L, et al. Maturational changes in the interdependencies between cortical brain areas of neonates during sleep. Cerebral Cortex. 2007; 17(3):583–90. https://doi.org/10.1093/cercor/bhk002 PMID: 16627860

48. Thiel M, Romano MC, Kurths J, Rolfs M, Kliegl R. Generating surrogates from recurrences. Philosophical Transactions of the Royal Society A. 2008; 366:545–57. https://doi.org/10.1098/rsta.2007.2109

49. Romano MC, Thiel M, Kurths J, Mergenthaler K, Engbert R. Hypothesis test for synchronization: twin surrogates revisited. Chaos. 2009; 19:015108. https://doi.org/10.1063/1.3072784 PMID: 19335012

50. Hegger R, Kantz H. Improved false nearest neighbor method to detect determinism in time series data. Physical Review E. 1999; 60(4):4970–3. https://doi.org/10.1103/PhysRevE.60.4970

51. Cabral J, Luckhoo H, Woolrich M, Joensson M, Mohseni H, Baker A, et al. Exploring mechanisms of spontaneous functional connectivity in MEG: How delayed network interactions lead to structured amplitude envelopes of band-pass filtered oscillations. NeuroImage. 2013; 90:423–35. https://doi.org/10.1016/j.neuroimage.2013.11.047 PMID: 24321555

52. Cooper GF. The computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence. 1990; 42(2-3):393–405. https://doi.org/10.1016/0004-3702(90)90060-D

53. Dagum P, Luby M. Approximating probabilistic inference in Bayesian belief networks is NP-hard. Artificial Intelligence. 1993; 60(1):141–153. https://doi.org/10.1016/0004-3702(93)90036-B

54. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning. 1992; 9(4):309–347. https://doi.org/10.1007/BF00994110

55. Buntine WL. A guide to the literature on learning probabilistic networks from data. Knowledge and Data Engineering, IEEE Transactions on. 1996 Apr; 8(2):195–210. https://doi.org/10.1109/69.494161

56. Heckerman D, Chickering DM. Learning Bayesian networks: The combination of knowledge and statistical data. In: Machine Learning; 1995. p. 20–197.

57. Allefeld C, Kurths J. An approach to multivariate phase synchronization analysis and its application to event-related potentials. International Journal of Bifurcation and Chaos. 2004; 14:417–26. https://doi.org/10.1142/S0218127404009521

58. Allefeld C, Muler M, Kurths J. Eigenvalue decomposition as a generalized synchronization cluster analysis. International Journal of Bifurcation and Chaos. 2007; 17(10):3493–7. https://doi.org/10.1142/S0218127407019251

59. Quinlan JR. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc.; 1993.

60. Hall MA. Correlation-based Feature Subset Seletion for Machine Learning; 1998.

61. Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human Brain Mapping. 2002; 15(1):1–25. https://doi.org/10.1002/hbm.1058 PMID: 11747097

62. Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature. 2006; 440(7087):1007–12. https://doi.org/10.1038/nature04701 PMID: 16625187

63. Al-Khassaweneh M, Villafane-Delgado M, Mutlu AY, Aviyente S. A Measure of Multivariate Phase Synchrony Using Hyperdimensional Geometry. IEEE Transactions on Signal Processing. 2016 jun; 64(11):2774–2787. https://doi.org/10.1109/TSP.2016.2529586

64. Porz S, Kiel M, Lehnertz K. Can spurious indications for phase synchronization due to superimposed signals be avoided? Chaos: An Interdisciplinary Journal of Nonlinear Science. 2014; 24(3):033112. https://doi.org/10.1063/1.4890568

65. Christodoulakis M, Hadjipapas A, Papathanasiou ES, Anastasiadou M, Papacostas SS, Mitsis GD. On the Effect of Volume Conduction on Graph Theoretic Measures of Brain Networks in Epilepsy. In: Neuromethods. Humana Press; 2014. p. 103–130.

66. Chella F, Marzetti L, Pizzella V, Zappasodi F, Nolte G. Third order spectral analysis robust to mixing artifacts for mapping cross-frequency interactions in EEG/MEG. NeuroImage. 2014; 91:146–61. https://doi.org/10.1016/j.neuroimage.2013.12.064 PMID: 24418509

67. Chella F, Pizzella V, Zappasodi F, Nolte G, Marzetti L. Bispectral pairwise interacting source analysis for identifying systems of cross-frequency interacting brain sources from electroencephalographic or magnetoencephalographic signals. Physical Review E. 2016; 93(5):052420. https://doi.org/10.1103/PhysRevE.93.052420 PMID: 27300936

68. Terrien J, Germain G, Marque C, Karlsson B. Bivariate piecewise stationary segmentation; improved pre-treatment for synchronization measures used on non-stationary biological signals. Medical Engineering & Physics. 2013; 35(8):1188–96. https://doi.org/10.1016/j.medengphy.2012.12.010

69. García-Prieto J, Bajo R, Pereda E. Efficient Computation of Functional Brain Networks: toward Real-Time Functional Connectivity. Frontiers in Neuroinformatics. 2017 feb; 11:8. https://doi.org/10.3389/fninf.2017.00008 PMID: 28220071

70. Gorgolewski KJ, Poldrack RA. A Practical Guide for Improving Transparency and Reproducibility in Neuroimaging Research. PLoS Biology. 2016 jul; 14(7):e1002506. https://doi.org/10.1371/journal.pbio.1002506 PMID: 27389358