Journal of Neuroscience Methods 250 (2015) 126–136

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational Neuroscience

Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy

Etienne Combrisson [a,b], Karim Jerbi [a,c,*]

[a] DYCOG Lab, Lyon Neuroscience Research Center, INSERM U1028, UMR 5292, University Lyon I, Lyon, France
[b] Center of Research and Innovation in Sport, Mental Processes and Motor Performance, University of Lyon I, Lyon, France
[c] Psychology Department, University of Montreal, QC, Canada

ARTICLE INFO

Article history:
Received 28 July 2014
Received in revised form 6 January 2015
Accepted 7 January 2015
Available online 14 January 2015

Keywords:
k-Fold cross-validation
Small sample size
Classification
Multi-class decoding
Brain–computer-interfaces (BCIs)
Machine learning
Binomial cumulative distribution
Classification significance
Decoding accuracy
MEG
ECoG
Intracranial EEG

ABSTRACT

Machine learning techniques are increasingly used in neuroscience to classify brain signals. Decoding performance is reflected by how much the classification results depart from the rate achieved by purely random classification. In a 2-class or 4-class classification problem, the chance levels are thus 50% or 25% respectively. However, such thresholds hold for an infinite number of data samples but not for small data sets. While this limitation is widely recognized in the machine learning field, it is unfortunately sometimes still overlooked or ignored in the emerging field of brain signal classification. Incidentally, this field is often faced with the difficulty of low sample size. In this study we demonstrate how applying signal classification to Gaussian random signals can yield decoding accuracies of up to 70% or higher in two-class decoding with small sample sets. Most importantly, we provide a thorough quantification of the severity and the parameters affecting this limitation using simulations in which we manipulate sample size, class number, cross-validation parameters (*k*-fold, leave-one-out and repetition number) and classifier type (Linear-Discriminant Analysis, Naïve Bayesian and Support Vector Machine). In addition to raising a red flag of caution, we illustrate the use of analytical and empirical solutions (binomial formula and permutation tests) that tackle the problem by providing statistical significance levels (*p*-values) for the decoding accuracy, taking sample size into account. Finally, we illustrate the relevance of our simulations and statistical tests on real brain data by assessing noise-level classifications in Magnetoencephalography (MEG) and intracranial EEG (iEEG) baseline recordings.

1. Introduction

Applying machine learning algorithms to brain signals in order to predict intentions or decode cognitive states has become an increasingly popular technique over the last decade. The surge in the use of machine learning methods in neuroscience has been largely fueled by the tremendous increase in brain–computer interface (BCI) and brain signal decoding research either using non-invasive recordings such as Electroencephalography (EEG) or Magnetoencephalography (MEG) (e.g. Aloise et al., 2012; Besserve et al., 2007; Jerbi et al., 2011; Krusienski and Wolpaw, 2009; Toppi et al., 2014; Waldert et al., 2008) or with intracranial EEG (e.g. Ball et al., 2009; Derix et al., 2012; Hamamé et al. (2012); Korczyn et al. (2013); Lachaux et al., 2007a,b; Leuthardt et al., 2004, 2006; Mehring et al., 2004; Pistohl et al., 2012; Schalk et al., 2008; Jerbi et al., 2007a,2009a,2013). Machine learning and signal classification techniques are powerful and complex tools that have to be used with caution. While most machine learning experts are well aware of the various caveats to watch out for, certain theoretical limitations of these methods can easily elude students and neuroscience researchers new to the field of machine learning and brain–computer interface research.

In supervised learning, samples of a subset of the data and knowledge of their corresponding class (label) are used to train a model to distinguish between two or more classes. The trained classifier is then tested on the remaining data samples (the hold-out samples). This procedure is generally repeated several times by varying the subsets used for training and those used for testing, a standard procedure known as cross-validation. The percent of over-all correct label (or class) prediction across the test samples of the multiple folds is known as the correct classification

* Corresponding author at: Psychology Department, University of Montreal, QC, Canada.
E-mail address: karim.jerbi@umontreal.ca (K. Jerbi).

http://dx.doi.org/10.1016/j.jneumeth.2015.01.010
0165-0270/

rate (sometimes called decoding accuracy). Conversely, the mean of misclassified samples over the folds is a measure of classifier prediction error.

The performance of a classifier in neural decoding studies is often assessed by how close its correct classification rate is to the maximum of 100%, or alternatively, how strongly it departs from the *chance-level* rate achieved by a classifier that would randomly associate the samples to the various classes. For instance, in a two-class or four-class classification problem, the probabilistic chance level indicating totally random classification is 50% or 25% respectively. Yet, although such probabilistic chance-levels widely applied in brain signal classification studies, they can be problematic because they are strictly speaking only valid for infinite sample sizes. While it will not come to anyone as a surprise that no study to date was able to acquire infinite data, it is intriguing how rarely brain signal classification studies acknowledge this limitation or take it into account. For a two-class classification problem with small sample size, 60%, 70% or even higher decoding percentages can in theory arise by chance (see simulation results below). As a consequence, for finite samples, a decoding percentage can only be considered reliable if it substantially, or better still, *significantly* departs from the theoretical level in statistical terms. But how can we assess the significance of the departure of a decoder from the outcome of total random classification? For a given sample size and a given number of classes, what would be the statistically significant threshold of correct classification that one needs to exceed in order to consider the decoding *statistically significant*? Although these questions have been widely recognized and addressed in the machine learning field (e.g. Kohavi, 1995; Martin and Hirschberg, 1996a,b), it is unfortunately often overlooked in the emerging field of brain signal classification which, incidentally, is often faced with low sample sizes for which the problem is even more critical.

Not all the previous brain decoding reports suffer from the caveat of using theoretical chance-level as reference. However, numerous studies only apply statistical assessment when testing for significant differences between the performance of multiple classifiers, or when comparing decoding across experimental conditions, but unfortunately neglect to provide a statistical assessment of decoding that accounts for sample size (e.g. Felton et al., 2007; Haynes et al., 2007; Bode and Haynes, 2009; Kellis et al., 2010; Hosseini et al., 2011; Sitaram et al., 2011; Hill et al., 2006; Wang et al., 2010; Bleichner et al., 2014; Babiloni et al., 2000; Ahn et al., 2013; Morash et al., 2008; Neuper et al., 2005; Kayikcioglu and Aydemir, 2010; Momennejad and Haynes, 2012). A number of such studies use theoretical percent chance-levels (e.g. 50% in a 2-class classification) as a reference against which classifier decoding performance is assessed. By doing so, such studies fail to account for the effect of finite sample size. This may have little effect in the case of large sample size or when extremely high decoding results are obtained, however, the bias and erroneous impact of such omissions can be critical for smaller sample sizes or when the decoding accuracies are barely above the theoretical chance levels.

Note however, that the rigorous assessment of significant classification thresholds is not equally ignored across the various types of neuronal decoding studies; it seems that the omissions (or unfortunate tendency to rely on the theoretical chance levels) are more common in more recent sub-branches of the neuronal decoding field. This is the case for signal classification and BCI studies based on non-invasive (fMRI, EEG and MEG) brain recordings in humans, and possibly electrocorticographic macro-electrode recordings in patients, where the methods (including classifiers, features and statistics) are less well-established than in the field of neuronal spike decoding in primates for instance.

In this brief article, we address caveats related to interpreting brain classification performances with small sample sizes. The paper is written with the broad neuroscience readership in mind and is oriented, in particular, to students and researchers new to neural signal classification. First of all, we describe how applying signal classification to randomly generated signals can yield decoding accuracies (correct classification rates) that strongly depart from theoretical chance levels, with values up to 70% and higher with small sample sizes (instead of the expected theoretical 50% for 2-class decoding). Most importantly, we illustrate and quantify the phenomenon by using simulations in which we manipulate sample size, class number, cross-validation parameters and classifier type. In addition to raising a red flag of caution, we recommend practical alternatives to overcome the problem. We describe a straight-forward method to derive a statistically significant threshold that accounts for sample size and provides confidence intervals for the classification accuracy achieved by cross-validations. A reference table is also provided to allow readers to quickly look-up the percent correct classification thresholds that need to be exceeded in order to assert statistical significance of the findings for a range of possible sample sizes, classes and significance levels.

2. Materials and methods

2.1. Data simulation and classification

2.1.1. Generating normally distributed random data

In order to simulate a situation with classification results that approach the theoretical chance level, we generated 100 data sets of zero-mean Gaussian white noise. The normally distributed variables in each data set were generated in MATLAB (Mathworks Inc., MA, USA) via a pseudo-random number generator. Each one of the 100 data sets was randomly split into c subsets data (here we used $c = 2$- or 4-classes) and we then evaluated the classification performance obtained by applying different classification algorithms to these simulated datasets. Because the variables in each 'simulated class' were drawn from the exact same Gaussian random distribution data set, applying supervised machine learning algorithms should fail to distinguish between classes and should theoretically yield chance-level classification rates (50% for $c = 2$ and 25% for $c = 4$). To examine the effect of sample size on how close the empirical classifications are to the theoretically expected chance level we varied the total number of samples n from 24 to 500. In other words, in the 2-class simulation for instance, the number of samples in each class varied from 12 to 250. Note that the code we implemented for the generation of random data for classification purposes is provided online (see Appendix A).

2.1.2. Classification algorithms

We implemented three types of machine learning algorithms: linear discriminant analysis (LDA), naïve Bayes (NB) classifier and a support vector machine (SVM), the latter with two different kernels: a linear kernel and a radial basis function (RBF) kernel. These three methods, which are frequently used for neural signal classification in the context of brain–computer interface research are briefly described in the following.

Linear discriminant analysis: LDA (Fisher, 1936) is a straightforward and fast algorithm which assumes that the independent variables in each class are normally distributed with identical covariance (homoscedasticity assumption). For a two dimension problem, the LDA tries to find a hyperplane that maximizes the mean distance between the two classes while minimizing the interclass variance. A multiclass problem can be tackled as a multiple two-class problem by discriminating each class from the rest using multiple hyperplanes.

Naive Bayesian classifier: The NB model (e.g. Fukunaga, 1990) is a probabilistic classifier that assigns features to the class to which they have the highest probability of belonging. NB assumes that the

features in each class are normally distributed and independent. The name arises from the fact that it is based on applying Bayes' theorem with strong (naive) independence assumptions.

Support vector machine: SVM (Boser et al. 1992; Burges, 1998; Cortes and Vapnik, 1995; Vapnik, 1995) classifiers originate from statistical learning theory. An SVM searches for a hyperplane that maximize margins between the hyperplane and the closest features in the training set. For non-linearly separable classes, SVM uses a kernel function to project features in a higher dimensional space in order to reduce the nonlinear problem to a linear one, which is then separable by a hyperplane. The (Gaussian) Radial Basis Function (RBF) kernel is a popular choice. In this study, both linear and RBF kernels were used for SVM classification.

Details of the theoretical background of various classifiers can be found in standard statistics and machine learning textbooks and various reviews (e.g. Lotte et al., 2007; Wieland and Pittore, 2014). Here, we used MATLAB implementation for the LDA and NB and the libsvm library for multi-class SVM.

2.1.3. *Repeated and stratified* k-*fold cross-validation*

To compute the decoding accuracy achieved by each one of the classifiers on the random data, we used standard stratified k-fold cross-validation. For a given data set size, all available N samples are partitioned into k folds, where $(k-1)$ folds are used for training the classifier model (training set) and the remaining fold is used for validation (test set). This procedure is then repeated k times so that each fold is used once as test set. The stratified option ensures that each fold has approximately the same proportion of samples from each class as in the original dataset as a whole. The case $k = N$ (e.g. 200 folds in a data set of 200 samples) is called leave-one-out (LOO) cross-validation because one element is used to test the performance of a classifier trained on the rest of the data. Because k-fold cross-validation involves a random partition, the variance of the classifier can in theory be reduced by repeating the full cross validation procedure q times. Therefore, in addition to testing different classifier types, this study explores the effect of the following parameters: n (sample size, 20–500), k (number of cross-validation folds: 5, 10 and leave-one-out) and q (number of repetitions: 1, 5 and 20).

2.2. *Statistical significance of classification using a binomial cumulative distribution*

For a given number of classes c, the percent theoretical chance level of classification is given by $100/c$. For example, for a 4-class problem, the chance level is $100/4 = 25\%$. This threshold is based on the assumption of infinite sample size. In practice, the empirical chance level depends on the number of samples available. One way to address this limitation is to test for the statistical significance of the decoding accuracy. This can be done by assuming that the classification errors obey a binomial cumulative distribution, where for a total of n samples and c classes, the probability to predict the correct class at least z times by chance is given by:

$$P(z) = \sum_{i=z}^{n} \binom{n}{i} \times \left(\frac{1}{c}\right)^{i} \times \left(\frac{c-1}{c}\right)^{n-1}$$

Although neural signal classification studies predominantly evaluate decoding performance by how well the results depart from the theoretical chance level, several BCI studies have in addition, used the binomial cumulative distribution to derive statistical significance thresholds (e.g. Ang et al., 2010; Demandt et al., 2012; Pistohl et al., 2012; Waldert et al., 2007, 2008, 2012). In this study, we use the MATLAB (Mathworks Inc., MA, USA) function *binoinv* to compute the statistically significant threshold $St(\alpha) = binoinv(1-\alpha, n, 1/c) \times 100/n$, where α is the significance level given by $\alpha = z/n$ (i.e. the ratio of tolerated false positives z – i.e. number of observations correctly classified by chance with respect to all observations n). For instance, for a sample size of $n = 40$ and a 2-class classification problem ($c = 2$), computing the threshold for statistical significance of the decoding at $\alpha = 0.001$ using the above formulation yields 70.0%. In other words, at $n = 40$, any decoding percentage below 70% is not statistically significant (at $p < 0.001$), whereas if one relied on the theoretical threshold for two classes (i.e. 50%) a decoding accuracy of 67% might have been considered relevant. Table 1 provides the minimal thresholds as a function of selected sample sizes, class number and significance levels. Note that code for the calculation of these analytical significance levels is provided online (see Appendix A).

2.3. *Statistical significance of classification using permutation tests*

The statistical significance of decoding can also be assessed by non-parametric statistical methods, namely using permutation tests (Good, 2000; Nichols & Holmes, 2002). By randomly permuting the observations across classes and calculating classification accuracy at each permutation, it is possible to establish an empirical null distribution of classification accuracies on random observations. The tails of this distribution can then be used to determine significance boundaries for a given rate of tolerated false positives (i.e. correct classifications that occur by chance). For instance, if the original (without randomization) classification accuracy is higher than the 95 percentile of empirical performance distribution established by randomly permuting the data, then one can assert that the original classification is significant with $p < 0.05$. The advantage of this empirical approach is that it does not require particular assumption about statistical properties of the samples.

An intuitive illustration of this procedure would be as follows: one performs for example 99 random permutations of the labels (classes) in the data and computes the classification accuracy for each permutation. This provides an empirical distribution of 99 classification accuracy values. Now if the classification performance obtained with the original (unpermuted) data is higher than the maximum of the empirical distribution, one can conclude that it is significant with $\alpha = 0.01$.

Permutations test provide a useful empirical approach to deriving statistical significance of classifier performance (e.g. Golland and Fischl, 2003; Ojala and Garriga, 2010; Meyers and Kreiman, 2011). To demonstrate the utility to derive significance boundaries as a function of sample size and thus compare it to the use of the binomial formula. To this end, we used simulated random data with associated labels (as described in Section 2.1) and computed the classification performance (using LDA) for 10,000 permutations (randomly exchanging labels of the original observations). From this we derived the accuracy thresholds that correspond to the 99%, 99.9% and 99.99% percentile of the distribution (i.e. $p < 0.01$, $p < 0.001$, and $p < 0.0001$ respectively). This was done for each sample size value n (20–500), which allowed us to depict the evolution of the empirical significance boundaries as a function of sample size. Note that code for the calculation of permutation-based empirical significance levels is provided online (see Appendix A).

2.4. *Classification of baseline data from real brain signals*

Because real data does not necessarily have the same properties as those implemented in our random data simulations (zero-mean Gaussian white noise), we also calculated the correct classification rate (as a function of sample size) that is achieved when classifying real brain data that do not contain any true discrepancies. This was carried out for pre-stimulus or baseline recordings in MEG (4 subjects) and with intracranial EEG recordings (4 patients). The

Table 1
Look-up table for statistically significant classification performance. Minimal correct classification rate (%) to assert statistical significance (at a given p-value) as a function of sample size n and number of classes c. Threshold values are based on the binomial cumulative distribution function and are rounded to the first digit.

n	c											
	2-Classes				4-Classes				8-Classes			
	$p<0.05$	$p<0.01$	$p<10^{-3}$	$p<10^{-4}$	$p<0.05$	$p<0.01$	$p<10^{-3}$	$p<10^{-4}$	$p<0.05$	$p<0.01$	$p<10^{-3}$	$p<10^{-4}$
20	70.0%	75.0%	85.0%	90.0%	40.0%	50.0%	55.0%	65.0%	25.0%	30.0%	40.0%	45.0%
40	62.5%	67.5%	75.0%	77.5%	37.5%	42.5%	47.5%	52.5%	22.5%	25.0%	30.0%	35.0%
60	60.0%	65.0%	70.0%	73.3%	35.0%	38.3%	43.3%	46.7%	20.0%	23.3%	26.7%	30.0%
80	58.7%	62.5%	67.5%	70.0%	32.5%	36.2%	41.2%	43.7%	18.7%	21.2%	25.0%	27.5%
100	58.0%	62.0%	65.0%	68.0%	32.0%	35.0%	39.0%	42.0%	18.0%	21.0%	24.0%	26.0%
200	56.0%	58.0%	61.0%	63.0%	30.0%	32.5%	35.0%	37.0%	16.5%	18.0%	20.0%	22.0%
300	54.7%	56.7%	59.0%	60.7%	29.0%	31.0%	33.0%	34.7%	15.7%	17.0%	18.7%	20.0%
400	54.0%	55.7%	57.7%	59.2%	28.5%	30.0%	31.7%	33.2%	15.2%	16.5%	17.7%	19.0%
500	53.6%	55.2%	57.0%	58.2%	28.2%	29.6%	31.2%	32.4%	15.0%	16.0%	17.2%	18.2%

rationale here is that baseline (pre-stimulus) data is not expected to show any genuine discriminative brain patterns related to post-stimulus events, and as such, it is comparable to random background noise. Therefore, signal classification on these baseline periods should fail, and the accuracies that classifiers achieve can be taken as an empirical representation of chance-level decoding.

2.4.1. Illustrative data from MEG rest activity

We used illustrative data from 4 subjects scanned with a whole-head MEG system (151 sensors; VSM MedTech, BC, Canada) acquired at 1250 Hz sampling rate and with a band pass filter of 0–200 Hz. The participants provided written informed consent, and the experimental procedures were approved by the Institutional Review Board and by the National French Science Ethical Committee. The MEG data segments used for the purpose of the current analysis were extracted from the pre-stimulus baseline of a visuomotor MEG experiment (Jerbi et al., 2007b), and each trial was assigned one of 2 (or of 4) arbitrary labels for the 2-class (or 4-class) classification. Oscillatory alpha (8–12 Hz) power was computed using Hilbert transform and subsequently used as feature in an LDA-based classification procedure. We used 10-fold cross-validation and the whole procedure was repeated for increasing values of trial numbers (sample size n) ranging from 20 to 200 (in steps of 8).

2.4.2. Illustrative data from intracranial EEG baseline activity

We used illustrative data from 4 epilepsy patients stereotactically implanted with intracranial depth electrodes (O.8 mm diameter, 10–15 contact leads, DIXI Medical Instruments, Besançon, France). The intracerebral EEG (iEEG) recordings were conducted using a video-SEEG monitoring system (Micromed, Treviso, Italy), which allowed for the simultaneous recording from 128 depth-EEG electrode sites (More details of the routine SEEG acquisitions in Jerbi et al., 2009b). The data were bandpass filtered online from 0.1 to 200 Hz and sampled at 1024 Hz. The recordings were performed at the epilepsy department of the Grenoble University Hospital (headed by Dr. Philippe Kahane). All participants provided written informed consent, and the experimental procedures were approved by the Institutional Review Board and by the National French Science Ethical Committee.

The data segments used here were extracted from the pre-stimulus (baseline) of a standard motor task and each trial was associated with one of 2 (or of 4) labels for the 2-class (or 4-class) classification. The labels assigned to each pre-stimulus baseline trial were in fact the genuine post-stimulus events for the same trials (but no true discrimination can be expected prior to stimulus onset as the post-stim event could not be known or inferred during the pre-stimulus period). Broadband gamma (60–250 Hz) power was computed using Hilbert transform and subsequently used as feature in an LDA-based classification procedure. As for the MEG data, we used 10-fold cross-validation and the whole procedure was repeated for increasing values of trial numbers (sample size n) ranging from 20 to 200 (in steps of 8).

3. Results

3.1. Empirical evaluation of chance level decoding as a function of sample size

Fig. 1 shows the decoding accuracies obtained by conducting 10-fold cross validation on 100 randomly generated data sets. The decoding is depicted as a function of increasing sample size (from 24 to 500) and for the case of 2-class (left column) and 4-class (right column) classification. Although the theoretical chance levels for these configurations are 50% and 25% respectively, the results show how much the empirical decoding accuracies obtained with random data deviate from these probabilistic values.

The small sample size problem: as expected, the variance of the decoding accuracy across the 100 simulated random data sets is high, and the more so for small sample sizes. As illustrated in Fig. 1, while the decoding does converge toward the theoretical chance level as the sample size increases, the values achieved with small sample size ($n<100$) can be disturbingly high. For instance, the highlighted examples (solid black line) in panels (a) to (f) illustrate how decoding accuracies as high as 70% for 2-class classification (or 50% for 4-classes) can be observed even when conducting classification on subsets of randomly generated data with randomly associated labels.

The small sample issue is persistent and qualitatively similar across all classifiers used. The first three rows of Fig. 1 show the results obtained with LDA, NB and SVM (with an RBF kernel). Panels (g) and (h) of Fig. 1 show that cross-validation results in all three classifiers have comparable deviation across the 100 simulated data sets. The variance of cross-validation over the 100 random data sets is high for small sample sizes (<200 observations) and drops off with increasing sample size.

3.2. Tweaking cross-validation parameters does not solve the small sample problem

It might be tempting to think that changing the cross-validation parameters might be a way to get around the small sample problem illustrated here. To address this we evaluated the impact of varying (a) the number of cross-validation folds, and (b) the number of repetitions of the cross-validation, on the reported deviation of the cross-validation results (cf. Fig. 1g and h) across the 100 data sets and all sample sizes. The results in Fig. 2(a–c) show that applying 5- and 10-fold cross-validation to the random data yielded substantially the same results, and that leave-one-out (LOO) cross-validation actually provided worse results (i.e.

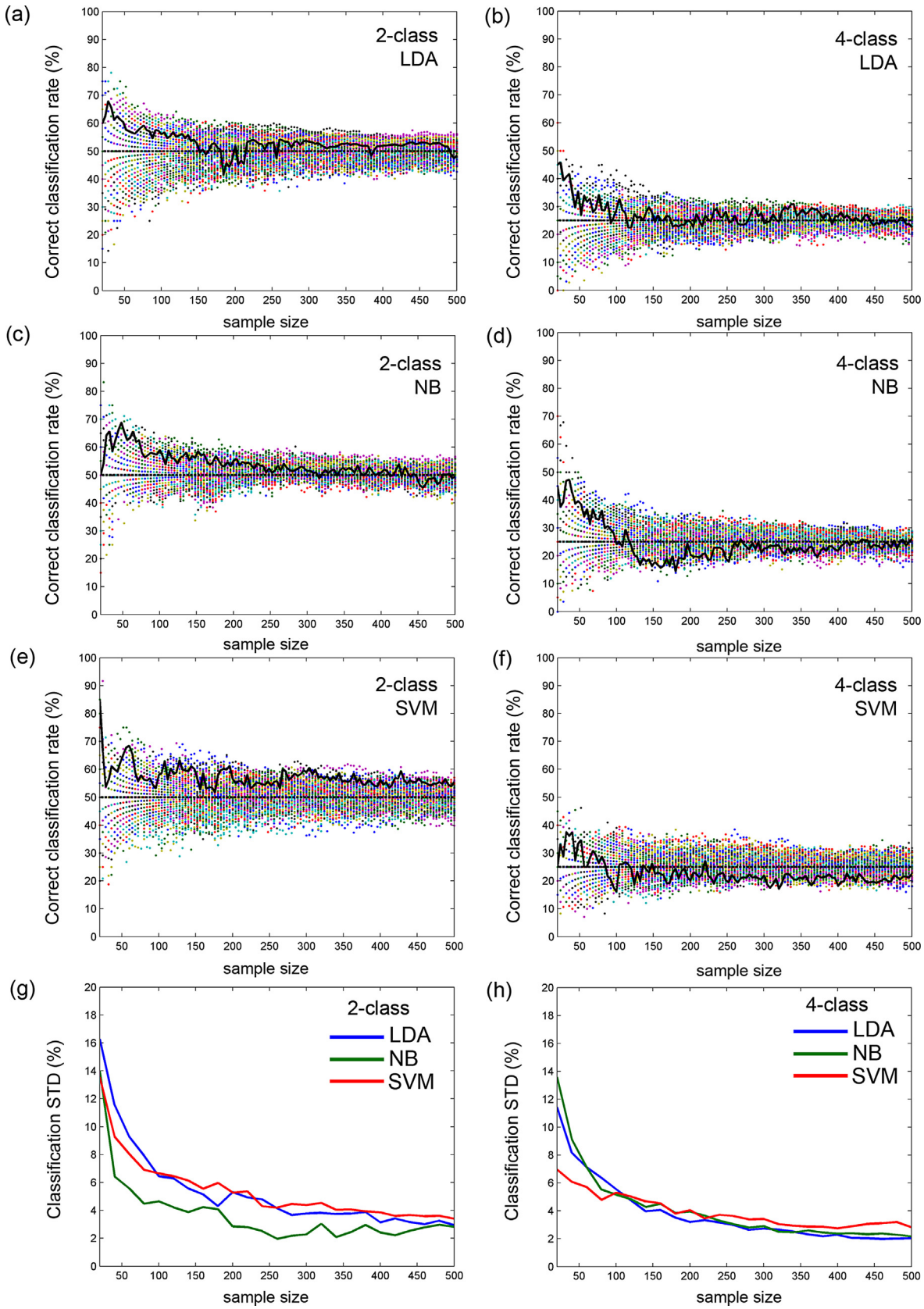

Fig. 1. Classifier decoding rates as a function of sample size when applied to random data sets using 10-fold cross-validation. (a) Two-class LDA classification rate (%) as a function of sample size (empirical results increasingly deviate from the 50% chance-level as the sample size gets smaller). The backline line shows the evolution of cross-validation results for one specific data set out of the 100 depicted in multiple colors. (b) Same as panel (a) but using 4-class classification, i.e. at each sample size *n*, the data is split into 4 virtual classes instead of two, (c and d) Same as (a and b) but for a Naïve Bayesian classifier. (e and f) Same as (a and b) but for an SVM classifier using an RBF kernel. (g) Evolution of cross-validation standard deviation across the 100 data sets for each of the three classifiers for 2-class decoding. (h) Same as panel (g) but for 4-class decoding.

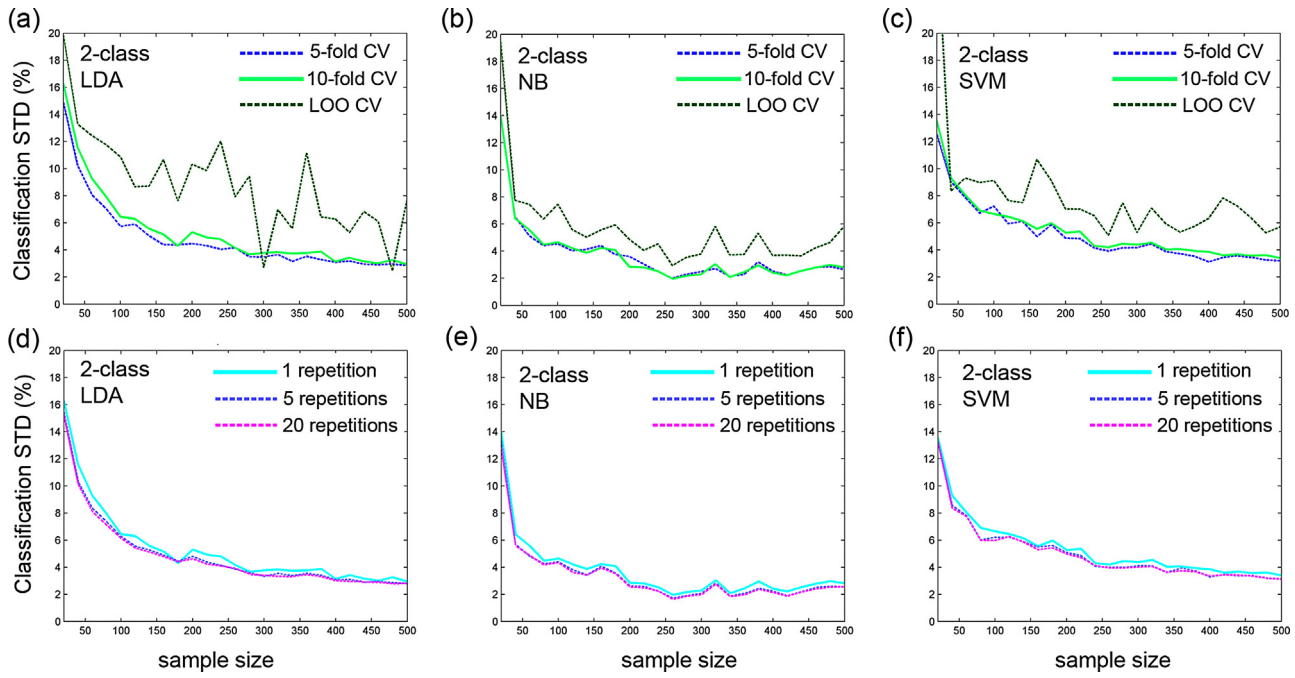

Fig. 2. Effect of cross-validation parameters on the variability of 2-class decoding performance computed across 100 sets of random data. (a–c) Effect of the number of folds (k): drop in cross-validation variance as sample size n increases, shown for $k = 5$, $k = 10$ (default), and $k = n$ (i.e. leave-one-out) and for all three classifiers LDA (panel a), NB (panel b) and SVM (panel c). (d–f) Effect of cross-validation repetition number: drop in cross-validation variance as sample size n increases, shown for repetition values $q = 1$ (default), $q = 5$, and $q = 20$ and for all three classifiers LDA (panel d), NB (panel e) and SVM (panel f). Note that the strong deviation from 50% chance-level for small sample sizes is persistent across all panels, and appears to be worst for LOO cross-validation with LDA.

higher variance). Moreover, repeating the cross-validation procedure (whether 5 or 20 times) achieved a negligible reduction of variance (Fig. 2d and e). Overall, these observations indicate that neither changing the number of folds nor to the number of overall repetitions has an impact on the variance of decoding accuracy (i.e. cross-validation results) across the 100 sets of Gaussian white noise.

3.3. Estimating statistical significance of decoding accuracy: binomial formula and permutation tests

Panels (a) and (b) in Fig. 3 show the evolution of the minimal statistically significant decoding rate as a function of sample size (respectively for 2- and 4-classes) using the binomial cumulative distribution (described in Section 2.2). The plots depicted for three distinct significance levels (10^{-2}, 10^{-3} and 10^{-4}) all show that the minimal correct decoding rate that is required in order to assert significance, decreases as the number of samples increases. Given small sample sizes (e.g. below 100 observations), to be statistically significant, the decoding accuracy must be substantially higher than the probabilistic chance level. For example, for 40 observations, a 2-class decoding is statistically significant (at $p < 0.001$) only if it exceeds the threshold of 75%. Note that for sample sizes as high as 500 observations, statistical significance still requires correct decoding higher than 55% (at $p < 0.01$), i.e. at least 5% above the theoretical chance level. A more comprehensive overview of the statistical decoding thresholds (wider ranges of p-values and of class number) computed for selected sample sizes (20–500), is provided in Table 1.

Panels (c) and (d) in Fig. 3 depict not only the evolution of the decoding boundaries for 2-class and 4-class decoding, using the binomial formula but also using the permutation test approach (see Section 2.3). Interestingly, the boundaries (for each level of admitted false positives) using both methods are reasonably close. The boundaries obtained with permutations show a slight tendency to be more restrictive than the binomial formula. While this is a little more apparent for small values of n, the difference between the two methods rapidly vanishes as n increases.

3.4. MEG and iEEG baseline data reveal erroneously high decoding results

Fig. 4 depicts the results of the empirical estimation of *de facto* chance-level decoding in illustrative MEG and iEEG data segments taken during pre-stimulus baseline periods (where no decoding is theoretically expected). Similarly to our findings using random data simulations (Fig. 1 a–f), the baseline MEG and iEEG data trials also led to decoding rates that strongly departed from the theoretical chance levels of 50% for 2-class classification and 25% for 4-class classification. Also in line with the results of the simulated data, the effects observed here were again highest for small sample sizes and dropped off slowly with increasing n. Note that the results in Fig. 4 show consistent performances across the 4 subjects at each value of n (with MEG and with iEEG). Finally, the superimposed gray curves (which depict the significance boundary given by the binomial formula as a function of sample size) nicely follow the trend of the % correct classification rate, and also illustrate cases of tolerated false positives for a given alpha.

4. Discussion

The current study has two primary take-home messages. The first is emphasizing the importance of watching out for a potential caveat that may arise when using departure from the theoretical chance-level as evidence for meaningful decoding. By launching various classifiers on normally distributed random data (Gaussian white noise), we demonstrate and quantify to which extent small samples lead to decoding accuracies that overshoot the chance-level merely by chance. This observation follows from the fact that

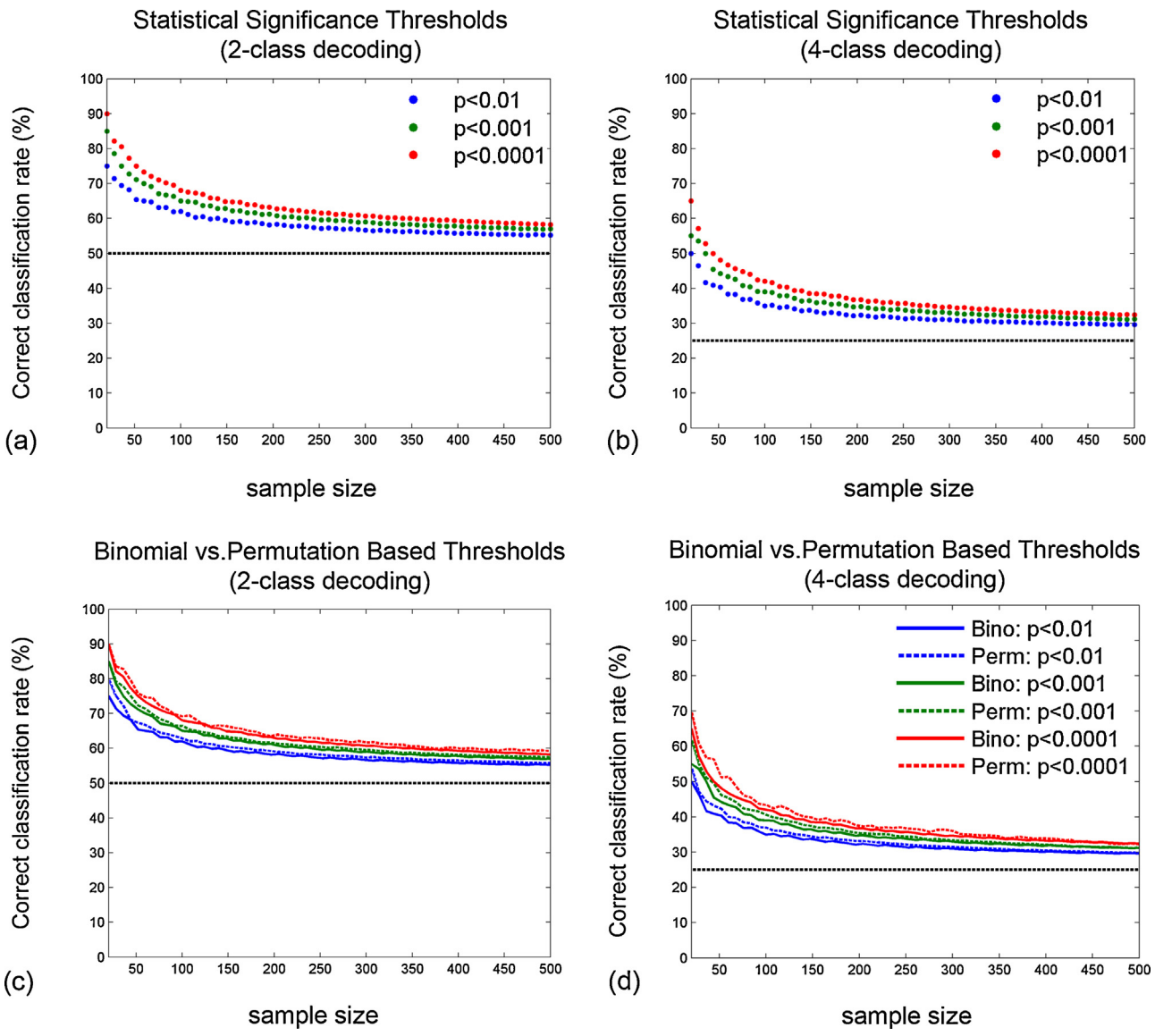

Fig. 3. Estimation of the statistical significance thresholds for 2- and 4-class classification as a function of sample size (assuming prediction errors are binomially distributed). Panels (a) and (b) show the evolution of the minimal statistically significant decoding rate as a function of sample size (respectively for 2 and 4 classes) using the binomial cumulative distribution (see Section 2.2). The plots were derived for significance levels 10^{-2}, 10^{-3} and 10^{-4}. As an example: panel (a) indicates that given a total of 100 data samples, a 2-class decoding result can only be considered statistically significant (at $p < 0.001$) if it exceeds 65%. This minimal value drops to 59% for 300 samples, but rises up to 75% if only 40 data points are available (See Table 1). Panels (c) and (d) show the same statistically significant decoding rate as a function of sample size (respectively for 2 and 4 classes) but now using both the binomial cumulative distribution (continuous lines) and the data-driven permutation-based approach (dashed lines) applied to the simulated random data (see Section 2.3 for details).

small samples are a bad approximation of true randomness and that as a result, the level 100/*c* (where c is the number of classes) is a purely theoretical chance-level that only holds for infinite sample sizes and that is particularly violated for small sample sizes. This basic fact is often overlooked in the neuronal decoding literature, where it is sometimes tempting to interpret for instance a 65% decoding accuracy in a 2-class classification as reflecting true neuronal decoding, without taking sample size into account. We have shown here that such levels of classification can be achieved with small samples of randomly generated data. This issue is not problematic for huge data samples, however, in data obtained from brain signal recordings in humans (such EEG or MEG), sample size can often be small. The effect of small samples on the reliability of probabilistic thresholds is therefore of particular importance in neural decoding and brain–computer interface studies. This effect is possibly even more critical when attempting to decode neuronal signals acquired using intracranial recordings (electrocorticography or stereoactic-EEG) and in clinical BCI applications where even less data samples might be available.

Furthermore, our exploration of the effects of classifier type (LDA, NB and SVM), cross-validation partition (number of folds) and cross-validation repetition number (up to 20), indicates that none of these parameters has a noticeable impact on the variance of the classification when applied to random data. The small sample size problem cannot be circumvented by tweaking these parameters and even for larger sample sizes of white noise any reduction in classification variance remains negligible. Note that the explored parameters and classifier comparisons performed here only address the variance and bias of the techniques when applied to normally distributed random data, reviews and comparisons of classifiers can be found elsewhere (e.g. Lotte et al., 2007).

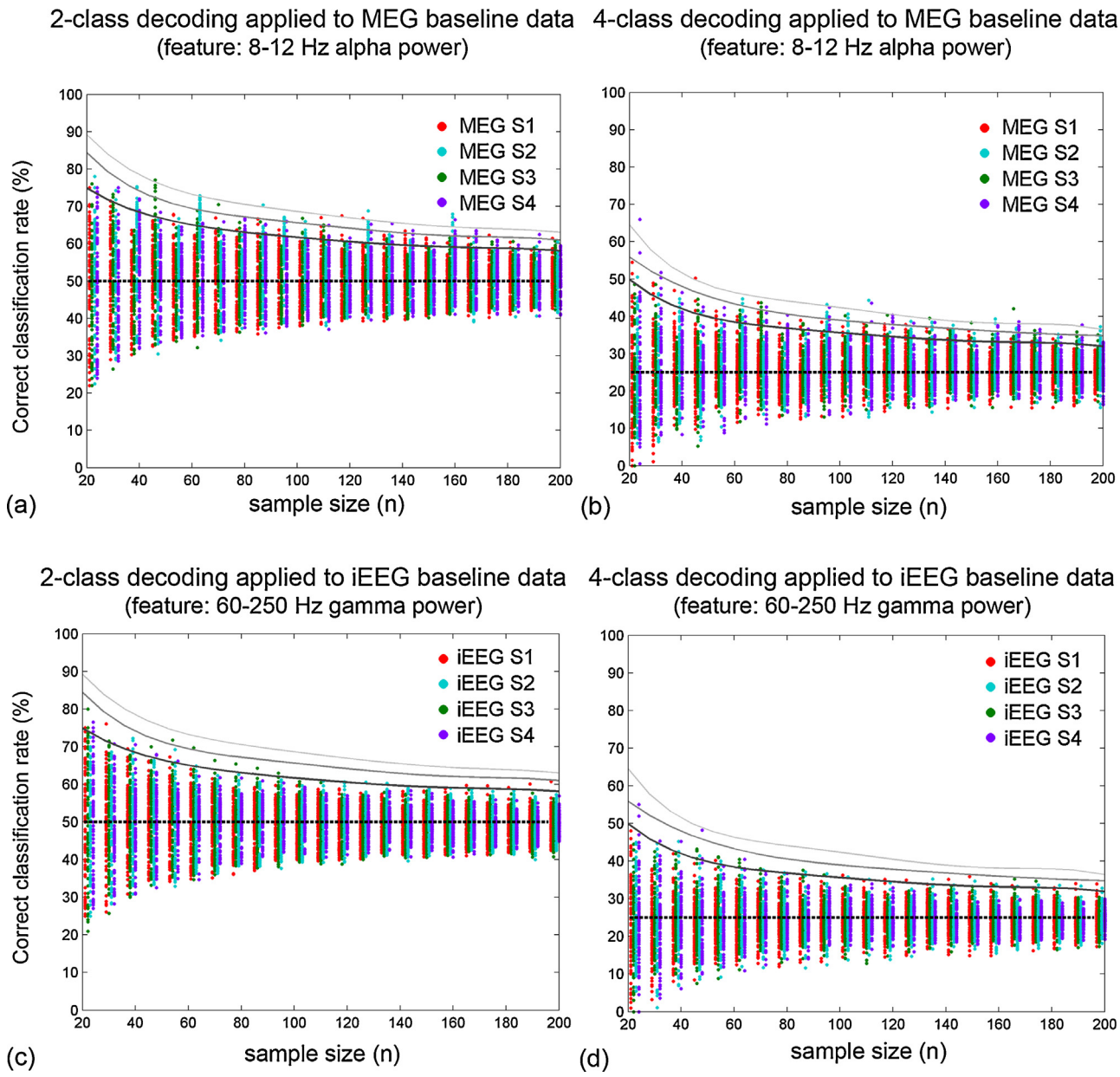

Fig. 4. Experimental assessment of chance-level classification accuracy in baseline (pre-stimulus) MEG and intracranial EEG data. (a) Two-class LDA classification rate (%) of MEG baseline data (alpha power features) as a function of sample size (illustrative data in 4 participants *MEG S1–S4*). The gray lines show the evolution of statistical significance boundaries computed with the binomial formula. Points lying above the gray lines thus represent false positives (type I errors) (b) Same as panel (a) but using 4-class classification, (c and d) Same as (a and b) but using baseline data (gamma power features) from intracranial EEG recordings (illustrative data in 4 epilepsy patients *iEEG S1–S4*).

Ten-fold cross-validation, which we used here as default, has been shown to be a reasonable choice providing low variance (Kohavi, 1995; Martin and Hirschberg, 1996a). Nevertheless, we also explored 5-fold and LOO cross-validation, alongside repetition number (Fig. 2). We found that none of these parameters could help reduce the cross-validation variance for low sample sizes. What is more, leave-one-out cross-validation showed even higher variability (in particular when using LDA), which is in agreement with previous reports suggesting that, despite its low bias, its high variance leads to unreliable estimates (Efron, 1983). Note that estimating the variance of cross-validation results across its k folds is generally problematic. Naive estimators that do not take into account error correlations due to the overlap between training and test sets (across the cross-validation folds) can severely underestimate variance (Bengio and Grandvalet, 2004). The cross-validation variances reported here were computed across the 100 independent data sets of Gaussian white noise.

The second take-home message from our study is an important reminder that one way to overcome this limitation is to seek statistically significant thresholds on decoding accuracy, rather than relying solely on the theoretical chance-level to claim successful decoding. This has been demonstrated here using a sample-size dependent threshold computation derived from the binomial cumulative distribution function. The underlying assumption that the number of errors is binomially distributed is commonly used in statistical learning (Kohavi, 1995; Breiman et al., 1984) but the statistical bounds it provides are unfortunately rarely exploited in brain signal decoding studies (e.g. Quiroga and Panzeri, 2009; Müller-Putz et al., 2008; Ang et al., 2010; Arvaneh et al., 2013; Demandt et al., 2012; Galan et al., 2014; Lampe et al., 2014; Pistohl

et al., 2012; Waldert et al., 2008, 2007, 2012). Kohavi (1995) provides a proof that *k*-fold cross-validation is binomial if the classifier induction method is stable under cross-validation. Note also that the validity of the assumption that prediction errors are binomially distributed has also been demonstrated for the specific case of 10-fold cross-validation with small samples (Martin and Hirschberg, 1996b). The latter study also emphasizes that the textbook formula based on the normal approximation to the binomial is not a good approximation to the confidence interval of an error rate estimate for small samples.

In addition to the binomial formula, we have also demonstrated the use of permutation tests as an alternative method to derive statistical significance boundaries for classifier performance as a function of sample size (Fig. 3c and d). Permutation tests provide a reliable and data-driven approach to the problem and has been proposed and used in numerous previous studies (e.g. Golland and Fischl, 2003; Ojala and Garriga, 2010; Meyers and Kreiman, 2011). Our analysis shows how, via multiple random shuffling of the data (or class labels), permutation tests can provide an estimate of sample-size dependent chance-level decoding accuracy. These empirical chance levels need to be exceeded in order to assert significance of a classification for a given rate of tolerated false positives. When applied to random noise signals, we found that the significance boundaries derived using permutations are reasonably close to those obtained using the binomial formula. Deciding which of the two approaches is more convenient when applied to real brain signals will likely depend on the data at hand. Permutation tests do not make any assumptions about the distribution of the data and provide a data-driven approach; however they also come with the burden of high computational cost, which dramatically increases with sample size, and with the level of statistical significance required.

Meyers and Kreiman (2011) note that deriving significance thresholds via the binomial formula as discussed here and elsewhere (e.g. Quiroga and Panzeri, 2009) comes with theoretical limitations that one should keep in mind, in particular, when combined with cross-validation; its application to mean performance over all folds violates the assumption of data point independence and leads to *p*-values that are too small. From a practical perspective, the impact of this theoretical limitation is likely to depend on the data at hand and on the selected cross-validation parameters. Simulations show that cross-validation parameters (number of cross-valdiation folds and repetitions) have an impact on the cumulative distribution function of classification accuracies (e.g. Noirhomme et al., 2014). As a result, cross-validation parameters, alongside classifier type and feature space, collectively lead to deviations from a binomial cumulative distribution. These deviations can be significant for small sample sizes (e.g. $N < 100$), which would advocate against using the binomial formula for statistical assessments under such circumstances (Noirhomme et al., 2014). In contrast, permutation tests being inherently data-driven, do take cross-validation parameters into account. As far as the Gaussian white noise simulated in the current study is concerned, permutation tests and the binomial formula appear to provide reasonably similar significance boundaries. Comparing the output of the binomial formula and (the more time consuming) permutation test, on at least a portion of the data, could be a pragmatic way to decide on whether the former provides a suitable and fast *approximation* of the latter.

Moreover, our analysis of decoding accuracy using real brain signals (with random labeling) is in line with our simulation results. This is a reassuring finding, as the latter were based on zero-mean Gaussian white noise while the former were based on power features (alpha and gamma-bands) derived from real brain data. The baseline-period MEG and iEEG data suggest that the binomial formula provides a reasonable estimation of chance-level decoding in these data sets. As a general rule, whenever possible, it is highly recommended to use baseline data as a recording in which no task-dependent encoding occurs and thus within chance-level decoding is expected. Comparisons with pre-stimulus (baseline) decoding performances should be used as an additional sanity check whenever such data is available (Meyers and Kreiman, 2011).

An alternative framework that can be applied to measure and compare classifier performance, is the use of receiver operating characteristic (ROC) analysis and in particular the area under the ROC curve (AUC) (Ling et al., 2003; Huang and Ling, 2005; Bradley, 1997). It has also been shown that calculating the probability density function (pdf) for each point on a ROC curve for any given sample size can be used to produce confidence intervals for ROC curves that are valid for small sample sizes (Tilbury et al., 2000). Adaptations of this method might be particularly suited to assessing classifier performance in BCI research (Hamadicharef, 2010). Other solutions that have been proposed to tackle the small sample size problem include frameworks that combine cross-validation with bootstrapping (e.g. Fu et al., 2005) and the use of class-dependent PCA in conjunction with linear discriminant feature extraction (Das and Nenadic, 2009). It is noteworthy that a few authors have even suggested that classification studies should be based primarily on effect size estimation with confidence intervals, rather than on significance tests and *p*-values (Berrar and Lozano, 2013).

In summary, the notion of statistical significance for decoding rates (or prediction error) and the small sample size problem have been tackled in the field of statistical learning for a long time (e.g. Raudys and Jain, 1991). However, these notions have not been sufficiently acknowledged in the relatively recent surge in application of machine learning methods in neuroscience. In the worse cases, this can unfortunately lead to erroneous interpretation of decoding results. Beyond its importance for brain–computer interface research specifically, signal classification is also increasingly used in neuroscience with the broader aim of elucidating the functional role of specific neuronal features (i.e. unraveling neuronal encoding by investigating single-trial neuronal decoding). Incidentally, this is where researchers are likely to be tempted to consider low (but above chance-level) decoding accuracies (e.g. 68% in a two-class classification) as being relevant. The use of confidence intervals and robust estimation of statistical significance is of particular importance in such studies, and even more so in cases with low trial numbers (e.g. below 150 observations). Machine learning and cross-validation accuracy in multi-class decoding may therefore not be thought of as a less-strict approach that can circumvent traditional rigorous statistical comparisons of data from multiple experimental conditions. Finally, whether signal classification is used in a BCI context *stricto sensu* or within a framework to conduct basic neuroscience analysis, we highly recommend systematically reporting the decoding accuracy as well as its statistical significance. We hope that the simulation results, statistical approaches and practical recommendations discussed here will be helpful in illustrating the problem and providing ways of tackling it.

Acknowledgements

Etienne Combrisson is currently supported by a Ph.D. Scholarship awarded by the Ecole Doctorale Inter-Disciplinaire Sciences-Santé (EDISS), Lyon, France. This work was partly performed within the framework of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program ANR-11-IDEX-0007. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program. The authors are grateful for the collaboration with the research and clinical staff of the Magnetoencephalography (MEG) center at the Pitié-Salpêtrière Hospital

in Paris and the University Hospital in Grenoble (Dr. Philippe Kahane).

Appendix A.

A. Software availability: The MATLAB scripts and functions that were developed and used in this study have been made available online for the community. The provided code can be used to generate, label and classify random data. It also provides routines to compute and plot, as a function of sample size, (a) analytical chance levels via the binomial formula as well as (b) empirical chance levels via permutation tests. We hope that this set of tools will help students and researchers replicate and extend our analyses. The code can be downloaded from Mathwork's File Exchange platform at the following URL: http://www.mathworks.fr/matlabcentral/fileexchange/48274-random-data-classification

References

Ahn M, Ahn S, Hong JH, Cho H, Kim K, Kim BS, et al. Gamma band activity associated with BCI performance: simultaneous MEG/EEG study. Front Hum Neurosci 2013:7, Available from: http://journal.frontiersin.org/Journal/10.3389/fnhum.2013.00848/full.

Aloise F, Schettini F, Aricò P, Salinari S, Babiloni F, Cincotti F. A comparison of classification techniques for a gaze-independent P300-based brain–computer interface. J Neural Eng 2012;9(4):045012.

Ang KK, Guan C, Sui Geok Chua K, Ang BT, Kuah C, Wang C, et al. Clinical study of neurorehabilitation in stroke using EEG-based motor imagery brain-computer interface with robotic feedback. IEEE 2010:5549–52 [cited 2014 Jul 4]. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5626782.

Arvaneh M, Guan C, Ang KK, Quek C. EEG data space adaptation to reduce intersession nonstationarity in brain–computer interface. Neural Comput 2013;25(May (8)):2146–71.

Babiloni F, Cincotti F, Lazzarini L, Millán J, Mouriño J, Varsta M, et al. Linear classification of low-resolution EEG patterns produced by imagined hand movements. IEEE Trans Rehabil Eng 2000;8(June (2)):186–8.

Ball T, Schulze-Bonhage A, Aertsen A, Mehring C. Differential representation of arm movement direction in relation to cortical anatomy and function. J Neural Eng 2009;6(1):016006.

Bengio Y, Grandvalet Y. No unbiased estimator of the variance of *k*-fold cross-validation. J Mach Learn Res 2004;5:1089–105.

Berrar D, Lozano JA. Significance tests or confidence intervals: which are preferable for the comparison of classifiers? J Exp Theor Artif Intell 2013;25(June (2)):189–206.

Besserve M, Jerbi K, Laurent F, Baillet S, Martinerie J, Garnero L, et al. Classification methods for ongoing EEG and MEG signals. Biol Res 2007;40(4):415–37.

Bleichner MG, Jansma JM, Sellmeijer J, Raemaekers M, Ramsey NF. Give me a sign: decoding complex coordinated hand movements using high-field fMRI. Brain Topogr 2014;27(March (2)):248–57.

Bode S, Haynes J-D. Decoding sequential stages of task preparation in the human brain. Neuroimage 2009;45(April (2)):606–13.

Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. ACM 1992:144–52 [cité 25.07.14].

Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30(July (7)):1145–59.

Breiman L, Friedman JH, Olshen R, Stone CJ. Classification and regression trees; 1984.

Burges CJ. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2(2):121–67.

Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(September (3)):273–97.

Das K, Nenadic Z. An efficient discriminant-based solution for small sample size problem. Pattern Recognit 2009;42(May (5)):857–66.

Demandt E, Mehring C, Vogt K, Schulze-Bonhage A, Aertsen A, Ball T. Reaching movement onset- and end-related characteristics of eeg spectral power modulations. Front Neurosci 2012:6, http://www.frontiersin.org/Journal/10.3389/fnins.2012.00065/full.

Derix J, Iljina O, Schulze-Bonhage A, Aertsen A, Ball T. "Doctor" or "darling"? Decoding the communication partner from ECoG of the anterior temporal lobe during non-experimental, real-life social interaction. Front Hum Neurosci 2012:6, http://www.frontiersin.org/Journal/10.3389/fnhum.2012.00251/full.

Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 1983;78(June (382)):316–31.

Felton EA, Wilson JA, Williams JC, Garell PC. Electrocorticographically controlled brain–computer interfaces using motor and sensory imagery in patients with temporary subdural electrode implants. Report of four cases. J Neurosurg 2007;106(March (3)):495–500.

Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugen 1936;7(September (2)):179–88.

Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics 2005;21(May (9)):1979–86.

Fukunaga K. Introduction to statistical pattern recognition. 2nd ed. Boston: Academic Press; 1990.

Galan F, Baker MR, Alter K, Baker SN. Missing kinaesthesia challenges precise naturalistic cortical prosthetic control, May. Report no.: 004861; 2014, http://biorxiv.org/lookup/doi/10.1101/004861.

Golland P, Fischl B. Permutation tests for classification: towards statistical significance in image-based studies. In: Proc. IPMI: international conference on information processing and medical imaging, LNCS, vol. 2732; 2003. p. 330–41.

Good PI. Permutation tests: a practical guide to resampling methods for testing hypotheses. 2nd ed. New York: Springer; 2000.

Hamadicharef B. AUC confidence bounds for performance evaluation of Brain-Computer Interface. In: IEEE 3rd International (Volume:5) Conference on Biomedical Engineering and Informatics (BMEI); 2010. p. 1988–91, Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5639671.

Hamamé CM, Vidal JR, Ossandón T, Jerbi K, Dalal SS, Minotti L, et al. Reading the mind's eye: online detection of visuo-spatial working memory and visual imagery in the inferior temporal lobe. NeuroImage 2012;59(January (1)):872–9.

Haynes J-D, Sakai K, Rees G, Gilbert S, Frith C, Passingham RE. Reading hidden intentions in the human brain. Curr Biol 2007;17(February (4)):323–8.

Hill NJ, Lal TN, Schröder M, Hinterberger T, Wilhelm B, Nijboer F, et al. Classifying EEG and ECoG signals without subject training for fast BCI implementation: comparison of nonparalyzed and completely paralyzed subjects. IEEE Trans Neural Syst Rehabil Eng 2006;14(June (2)):183–6.

Hosseini SMH, Mano Y, Rostami M, Takahashi M, Sugiura M, Kawashima R. Decoding what one likes or dislikes from single-trial fNIRS measurements. NeuroReport 2011;22(April (6)):269–73.

Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17(3):299–310.

Jerbi K, Bertrand O, Schoendorff B, Hoffmann D, Minotti L, Kahane P, et al. Online detection of gamma oscillations in ongoing intracerebral recordings: From functional mapping to brain computer interfaces. In: Noninvasive Funct Source Imaging Brain Heart Int Conf Funct Biomed Imaging 2007 NFSI-ICFBI 2007 Jt Meet 6th Int Symp On. IEEE; 2007a. p. 330–3, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4387767.

Jerbi K, Lachaux J-P, Karim N, Pantazis D, Leahy RM, Garnero L, et al. Coherent neural representation of hand speed in humans revealed by MEG imaging. Proc Natl Acad Sci 2007b;104(18):7676–81.

Jerbi K, Freyermuth S, Minotti L, Kahane P, Berthoz A, Lachaux J. Watching brain TV and playing brain ball International review of neurobiology. In: Brain machine interfaces for space applications: enhancing astronaut capabilities. San Diego: Elsevier Academic Press; 2009a. p. 159–68 [chapter 12]. http://linkinghub.elsevier.com/retrieve/pii/S0074774209860121.

Jerbi K, Ossandón T, Hamamé CM, Senova S, Dalal SS, Jung J, et al. Task-related gamma-band dynamics from an intracerebral perspective: review and implications for surface EEG and MEG. Hum Brain Mapp 2009b;30(June (6)):1758–71.

Jerbi K, Vidal JR, Mattout J, Maby E, Lecaignard F, Ossandon T, et al. Inferring hand movement kinematics from MEG, EEG and intracranial EEG: from brain-machine interfaces to motor rehabilitation. IRBM 2011;32(February (1)):8–18.

Jerbi K, Combrisson E, Dalal SS, Vidal JR, Hamame CM, Bertrand O, et al. Decoding cognitive states and motor intentions from intracranial EEG: how promising is high-frequency brain activity for brain–machine interfaces? Epilepsy Behav 2013;28(2):283–302.

Kayikcioglu T, Aydemir O. A polynomial fitting and k-NN based approach for improving classification of motor imagery BCI data. Pattern Recognit Lett 2010;31(August (11)):1207–15.

Kellis S, Miller K, Thomson K, Brown R, House P, Greger B. Decoding spoken words using local field potentials recorded from the cortical surface. J Neural Eng 2010;7(October (5)):056007.

Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection; 1995. p. 1137–45, http://frostiebek.free.fr/docs/Machine%20Learning/validation-1.pdf.

Korczyn AD, Schachter SC, Brodie MJ, Dalal SS, Engel J, Guekht A, et al. Epilepsy, cognition, and neuropsychiatry (Epilepsy, Brain, and Mind, part 2). Epilepsy Behav 2013;28(2):283–302.

Krusienski DJ, Wolpaw JR. Brain-computer interface research at the wadsworth center developments in noninvasive communication and control. Int Rev Neurobiol 2009;86:147–57.

Lachaux J-P, Jerbi K, Bertrand O, Minotti L, Hoffmann D, Schoendorff B, et al. A Blueprint for Real-Time Functional Mapping via Human Intracranial Recordings. PLoS ONE 2007a;2(October (10)):e1094.

Lachaux J-P, Jerbi K, Bertrand O, Minotti L, Hoffmann D, Schoendorff B, et al. BrainTV: a novel approach for online mapping of human brain functions. Biol Res 2007b;40(January (4)):401–13.

Lampe T, Fiederer LDJ, Voelker M, Knorr A, Riedmiller M, Ball T. A brain–computer interface for high-level remote control of an autonomous, reinforcement-learning-based robotic system for reaching and grasping. In: Proceedings of the 19th international conference on intelligent user interfaces. New York, NY, USA: ACM; 2014. p. 83–8, http://dx.doi.org/10.1145/2557500.2557533.

Leuthardt EC, Schalk G, Wolpaw JR, Ojemann JG, Moran DW. A brain–computer interface using electrocorticographic signals in humans. J Neural Eng 2004;1(June (2)):63.

Leuthardt EC, Miller KJ, Schalk G, Rao RPN, Ojemann JG. Electrocorticography-based brain computer interface–the Seattle experience. IEEE Trans Neural Syst Rehabil Eng 2006;14(June (2)):194–8.

Ling CX, Huang J, Zhang H. AUC: a statistically consistent and more discriminating measure than accuracy; 2003. p. 519–24, http://arion.csd.uwo.ca/faculty/ling/papers/ijcai03.pdf.

Lotte F, Congedo M, Lécuyer A, Lamarche F, Arnaldi B. A review of classification algorithms for EEG-based brain–computer interfaces. J Neural Eng [Internet] 2007 [cited 2012 Oct 3];4. Available from: http://hal.archives-ouvertes.fr/docs/00/13/49/50/PDF/article.pdf.

Martin JK, Hirschberg DS. Small sample statistics for classification error rates I: error rate measurements. Irvine: Information and Computer Science, University of California; 1996a, http://www.ics.uci.edu/~dan/pubs/TR96-21.pdf.

Martin JK, Hirschberg DS. Small sample statistics for classification error rates II: confidence intervals and significance tests [Internet]. Information and Computer Science. Irvine: University of California; 1996b, Disponible sur: http://www.ics.uci.edu/~dan/pubs/TR96-22.pdf.

Mehring C, Nawrot MP, de Oliveira SC, Vaadia E, Schulze-Bonhage A, Aertsen A, et al. Comparing information about arm movement direction in single channels of local and epicortical field potentials from monkey and human motor cortex. J Physiol – Paris 2004;98(July (4–6)):498–506.

Meyers EM, Kreiman G. Tutorial on pattern classification in cell recordings. In: Kriegeskorte N, Kreiman G, editors. Understanding visual population codes. Boston: MIT Press; 2011.

Momennejad I, Haynes J-D. Human anterior prefrontal cortex encodes the "what" and "when" of future intentions. Neuroimage 2012;61(May (1)):139–48.

Morash V, Bai O, Furlani S, Lin P, Hallett M. Classifying EEG signals preceding right hand, left hand, tongue, and right foot movements and motor imageries. Clin Neurophysiol 2008;119(November (11)):2570–8.

Müller-Putz GR, Scherer R, Brunner C, Leeb R, Pfurtscheller G. Better than random? A closer look on BCI results. Int J Bioelectromagn 2008;10(1):52–5.

Neuper C, Scherer R, Reiner M, Pfurtscheller G. Imagery of motor actions: differential effects of kinesthetic and visual-motor mode of imagery in single-trial EEG. Brain Res Cogn Brain Res 2005;25(December (3)):668–77.

Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 2002;15(1):1–25.

Noirhomme Q, Lesenfants D, Gomez F, Soddu A, Schrouff J, Garraux G, et al. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. NeuroImage: Clin 2014;4:687–94.

Ojala M, Garriga GC. Permutation tests for studying classifier performance. J Mach Learn Res 2010;11(June):1833–63.

Pistohl T, Schulze-Bonhage A, Aertsen A, Mehring C, Ball T. Decoding natural grasp types from human ECoG. NeuroImage 2012;59(January (1)):248–60.

Quiroga RQ, Panzeri S. Extracting information from neuronal populations: information theory and decoding approaches. Nat Rev Neurosci 2009;10:173.

Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 1991;13(March (3)):252–64.

Schalk G, Miller KJ, Anderson NR, Wilson JA, Smyth MD, Ojemann JG, et al. Two-dimensional movement control using electrocorticographic signals in humans. J Neural Eng 2008;5(March (1)):75–84.

Sitaram R, Lee S, Ruiz S, Rana M, Veit R, Birbaumer N. Real-time support vector classification and feedback of multiple emotional brain states. NeuroImage 2011;56(May (2)):753–65.

Tilbury JB, Van Eetvelt WJ, Garibaldi JM, Curnsw JSH, Ifeachor EC. Receiver operating characteristic analysis for intelligent medical systems-a new approach for finding confidence intervals. IEEE Trans Biomed Eng 2000;47(7):952–63.

Toppi J, Risetti M, Quitadamo LR, Petti M, Bianchi L, Salinari S, et al. Investigating the effects of a sensorimotor rhythm-based BCI training on the cortical activity elicited by mental imagery. J Neural Eng 2014;11(June (3)):035010.

Vapnik V. The nature of statistical learning theory. New York: Springer Science & Business Media; 1995.

Waldert S, Braun C, Preissl H, Birbaumer N, Aertsen A, Mehring C. Decoding performance for hand movements: EEG vs. MEG. IEEE 2007:5346–8.

Waldert S, Preissl H, Demandt E, Braun C, Birbaumer N, Aertsen A, et al. Hand movement direction decoded from MEG and EEG. J Neurosci 2008;28(January (4)):1000–8.

Waldert S, Tüshaus L, Kaller CP, Aertsen A, Mehring C. fNIRS exhibits weak tuning to hand movement direction. PLoS ONE 2012;7(November (11)):e49266.

Wang W, Sudre GP, Xu Y, Kass RE, Collinger JL, Degenhart AD, et al. Decoding and cortical source localization for intended movement direction with MEG. J Neurophysiol 2010;104(November (5)):2451–61.

Wieland M, Pittore M. Performance Evaluation of Machine Learning Algorithms for Urban Pattern Recognition from Multi-spectral Satellite Images. Remote Sens 2014;6(March (4)):2912–39.