Chapter 17
Is It Significant? Guidelines for Reporting BCI Performance

Martin Billinger, Ian Daly, Vera Kaiser, Jing Jin, Brendan Z. Allison, Gernot R. Müller-Putz, and Clemens Brunner

Abstract Recent growth in brain-computer interface (BCI) research has increased pressure to report improved performance. However, different research groups report performance in different ways. Hence, it is essential that evaluation procedures are valid and reported in sufficient detail.

In this chapter we give an overview of available performance measures such as classification accuracy, cohen's kappa, information transfer rate (ITR), and written symbol rate (WSR). We show how to distinguish results from chance level using confidence intervals for accuracy or kappa. Furthermore, we point out common pitfalls when moving from offline to online analysis and provide a guide on how to conduct statistical tests on BCI results.

17.1 Introduction

Brain–computer interface (BCI) research is expanding in many ways. Within the academic research community, new articles, events, and research groups emerge increasingly quickly. Research labs have developed BCIs for communication [7, 19, 27, 34, 43–45, 67], for control of wheelchairs [24, 53] and neuroprosthetic devices

M. Billinger (✉) · I. Daly · V. Kaiser · B.Z. Allison · G.R.Müller-Putz · C. Brunner
Institute for Knowledge Discovery, Graz University of Technology, Austria
e-mail: martin.billinger@tugraz.at; ian.daly@tugraz.at; vera.kaiser@tugraz.at; allison@tugraz.at; gernot.mueller@tugraz.at; clemens.brunner@tugraz.at

C. Brunner
Swartz Center for Computational Neuroscience, INC, UCSD, San Diego, CA, USA
e-mail: clbrunner@ucsd.edu

J. Jin
Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, China
e-mail: jinjing@ecust.edu.cn

B. Z. Allison et al. (eds.), *Towards Practical Brain-Computer Interfaces*, Biological and Medical Physics, Biomedical Engineering, DOI 10.1007/978-3-642-29746-5_17,

[31, 46]. Although BCI research has been conducted for more than 20 years now, only some research labs have successfully applied BCIs to patient use [30, 36–38, 47, 49, 51, 52, 65]. The popular media has also shown increased interest in BCIs, with BCIs featured prominently in science fiction as well as in the mainstream. Additionally, new businesses are gaining attention with various products sold as BCIs for entertainment.

Hence, there is growing attention in performance, and increased pressure to report improved performance. Recent articles that developed fast BCIs openly noted this feat [6, 12, 63]. Articles routinely highlight methods and results that improve accuracy or reduce illiteracy relative to earlier work [3, 4, 10, 11, 33, 55, 62]. However, different groups use different methods for reporting performance, and it is essential that (1) the evaluation procedure is valid from a statistical and machine learning point of view, and (2) this procedure is described in sufficient detail.

It is also important to distinguish any reported BCI performance from the chance level, the expected best performance obtainable by chance alone. Depending on the performance measure, the number of classes in the BCI task, and the number of available trials, the chance level varies and should be considered in every study [48].

In this chapter, we provide an introduction to common performance measures (such as classification accuracy, Cohen's kappa, and information transfer rate). Furthermore, we discuss confidence intervals of the classification accuracy and Cohen's kappa to estimate the associated chance level. We also summarize state of the art offline procedures to estimate performance on a pre-recorded data set and discuss common cross-validation pitfalls. In the last two sections, we describe statistical tests often used in BCI studies, such as t-tests, repeated measures ANOVA, and suitable post-hoc tests. We also mention the need to correct for multiple comparisons.

17.2 Performance Measures

17.2.1 Confusion Matrix

A number of metrics may be used to measure the performance of a BCI. These include the number of correct classifications and the number of mistakes made by the classifier. The most straightforward classification example is binary classification, in which the classifier need only differentiate two classes. For example, this might be the case in the popular P300 speller first presented by Farwell and Donchin [22]. The task of the classifier is to determine if there is a P300 event present in a particular time segment of the EEG. Therefore, the two classes are either "yes, there is a P300 present" or "no, there is no P300 present." When considering such binary classification problems, four classification results are possible:

(1) A trial is classified as containing a P300 when a P300 is present (true positive, TP).

Table 17.1 Confusion matrix for binary classification

		Predicted		
		Class 1	Class 2	
Actual	Class 1	TP	FN	TP+FN
	Class 2	FP	TN	FP+TN
		TP+FP	FN+TN	N

Table 17.2 Example of a confusion matrix for three classes. The diagonal contains all 257 correct classifications (86 + 45 + 126), whereas the 193 misclassifications are on the off-diagonal (45 + 19 + 32 + 73 + 10 + 14). The sum of all elements yields 450 and equals the total number of trials (shown in the lower right corner). The row sums reflect the relative frequencies of each class (rightmost column). In this example, the classes are balanced, because each class occurs 150 times. The column sums reveal how many trials were classified as the specific class. In this example, the classifier assigned 169 trials to left hand, 104 trials to right hand, and 177 trials to foot imagery. Since these numbers are not equal, and because the classes were equally distributed, the classifier is biased towards left hand and foot classes

		Predicted			
		Left hand	Right hand	Foot	
Actual	Left hand	86	45	19	150
	Right hand	73	45	32	150
	Foot	10	14	126	150
		169	104	177	450

(2) A trial is classified as containing a P300 when a P300 is not present (false positive, FP).

(3) A trial is classified as not containing a P300 when there is no P300 present (true negative, TN).

(4) A trial is classified as not containing a P300 when there is a P300 present (false negative, FN).

For two or more classes, it is useful to employ a so-called confusion matrix to present the results. A confusion matrix presents the results of the classifier over several trials against the actual known classes of items in the dataset. This allows for an evaluation of which classes are being correctly and incorrectly classified. For binary classification described above, the structure of the confusion matrix is illustrated in Table 17.1.

Consider the case of a motor imagery based BCI with three possible classes. The BCI user may imagine left hand movement, right hand movement or foot movement to control the BCI. In the classification example illustrated in Table 17.2, 450 trials were classified into three different possible classes. The columns list the output from the classifier, while the rows list the actual class that the trials corresponded to. For example, 86 trials were correctly classified as corresponding to left hand imagery, whereas 45 trials that corresponded to left hand imagery were misclassified as corresponding to right hand imagery. From this example, it is clear that the number of correct classifications for each class are found along the diagonal of the confusion matrix. The row sums reflect the a priori distribution of the classes, that is, the relative frequency of each class. Conversely, the column sums reveal a potential bias of the classifier towards one (or more) classes.

While the confusion matrix contains all information on the outcome of a classification procedure, it is difficult to compare two or more confusion matrices. Therefore, most studies usually report scalar performance measures, which can be derived from the confusion matrix. Metrics commonly used in reporting BCI results include classification accuracy, Cohen's kappa κ, sensitivity and specificity, positive and negative predictive value, the F-measure and the r^2 correlation coefficient [57].

17.2.2 Accuracy and Error Rate

The accuracy p is the probability of performing a correct classification. It can be estimated from dividing the number of correct classifications by the total number of trials

$$p = \frac{\sum C_{i,i}}{N}. \tag{17.1}$$

$C_{i,i}$ is the ith diagonal element of the confusion matrix, and N is the total number of trials. The error rate or misclassification rate $e = 1 - p$ is the probability of making an incorrect classification.

Accuracy and error rate do not take class balance into account. If one class occurs more frequently than the other, accuracy may be high even for classifiers that cannot discriminate between classes. See Tables 17.3 and 17.4 for examples.

17.2.3 Cohen's Kappa

Cohen's kappa (κ) is a measure for the agreement between nominal scales [15]. As such κ can be used to measure the agreement between true class labels and classifier output. It is scaled between 1 (perfect agreement) and 0 (pure chance agreement). Equation (17.2) shows how to obtain κ from accuracy p and chance level p_0.

$$\kappa = \frac{p - p_0}{1 - p_0} \tag{17.2}$$

The chance level p_0 is the accuracy under the assumption that all agreement occurred by chance (see Sect. 17.3.1). p_0 can be estimated from the confusion matrix by

$$p_0 = \frac{\sum C_{i,:} C_{:,i}}{N^2}. \tag{17.3}$$

Table 17.3 Confusion matrix for binary classification when the two classes are not balanced (class 1 occurs more often than class 2). *Left*: The classifier selected the classes with a probability of 50 %. *Right*: The classifier always selected the first class

		Predicted		
		1	2	
Actual	1	45	45	90
	2	5	5	10
		50	50	100

$p = 0.5 \qquad \kappa = 0$

		Predicted		
		1	2	
Actual	1	90	0	90
	2	10	0	10
		100	0	100

$p = 0.9 \qquad \kappa = 0$

Table 17.4 Confusion matrix for binary classification when the two classes are not balanced (class 1 occurs more often than class 2). *Left*: The classifier selected the first class with 90 % probability and the second class with 10 % probability. *Right*: The classifier classified all trials correctly

		Predicted		
		1	2	
Actual	1	81	9	90
	2	9	1	10
		90	10	100

$p = 0.82 \qquad \kappa = 0$

		Predicted		
		1	2	
Actual	1	90	0	90
	2	0	10	10
		90	10	100

$p = 1 \qquad \kappa = 1$

$C_{i,:}$ and $C_{:,i}$ are the ith row and column of the confusion matrix, and N is the total number of trials.

For both confusion matrices in Table 17.3 $\kappa = 0$, indicating classification at chance level. Neither of these confusion matrices represents a meaningful classifier, although accuracies are 0.5 and 0.9 respectively.

17.2.4 Sensitivity and Specificity

Alternative metrics reported in BCI studies include the sensitivity and specificity (see for example [5,21,25,60]), which measure the proportion of correctly identified positive results (true positives) and the proportion of correctly identified negative results (true negatives). Sensitivity is defined as

$$H = \mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{17.4}$$

Specificity is then defined as

$$\mathrm{Sc} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{17.5}$$

The sensitivity is alternately referred to as the true positive rate (TPR) or the recall. The false positive rate (FPR) is then equal to 1—specificity.

The false detection rate F may be calculated as

$$F = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}. \tag{17.6}$$

From this, the positive predictive value (also referred to as the precision) may be calculated as $1 - F$. The HF difference $(H - F)$, as developed in [32], may then be derived.

These metrics may also be used to measure the receiver operator characteristic (ROC) curve [18, 29, 41]. This is a plot of how the true positive rate varies against the false positive rate for a binary classifier as the classification threshold is varied between its smallest and largest limit. The x axis of the ROC curve is the false positive rate (1—specificity), while the y axis is the true positive rate (sensitivity). The larger the area under the ROC curve, the larger the true positive rate and the smaller the false positive rate for a greater number of threshold values. Thus, an ROC curve that forms a diagonal from the bottom left corner of the plot to the top right is at theoretical chance level, whereas an ROC plot that reaches the top left corner is reporting perfect classification.

17.2.5 F-Measure

The terms precision and recall (sensitivity) may be used to describe the accuracy of classification results. Precision (also referred to as the positive predictive value) measures the fraction of classifications which are correct while recall measures the fraction of true positive classifications.

As the precision is increased, the recall decreases, and vice-versa. Therefore, for a given classifier, it is useful to have a measure of the harmonic mean of both measures. The F-measure is used to do this and is defined as

$$F_\alpha = \frac{(1 + \alpha) \cdot (1 - F) \cdot H}{\alpha \cdot (1 - F) + H}, \tag{17.7}$$

where α is the significance level of the measure and may be varied between 0 and 1. Thus, the F-measure may be analogous to the ROC curve, in that it provides a measure of the classifier performance across different significance levels.

17.2.6 Correlation Coefficient

The correlation coefficient may be used for either feature extraction or validation of classification results (see for example [13, 41, 60]). It is defined—via Pearson's correlation coefficient—as

$$r = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{(\sum_i (y_i - \bar{y})^2)(\sum_i (x_i - \bar{x})^2)}}, \tag{17.8}$$

where x_i denotes output values, y_i the class labels, $\bar{x}$ the mean of x and $\bar{y}$ the mean of the labels y.

Pearson's correlation should be used for Gaussian data, while for non-Gaussian data the rank correlation is recommended. The rank correlation is defined as above with the difference that x_i and y_i values are replaced by rank(x_i) and rank(y_i).

The correlation varies between -1 and 1, with a 0 indicating no correlation between the classifier results and a 1 indicating perfect positive correlation. A correlation of -1 indicates perfect negative correlation and may be discounted if the squared correlation measure is chosen (as used in [13]).

17.3 Significance of Classification

Reporting classification results by providing performance measures alone is often not enough. Even accuracies as high as 90 % can be meaningless if the number of trials is too low or classes are not balanced (see Table 17.3).

The practical level of chance [48] provides a convenient tool to quickly verify if an accuracy value lies significantly above chance level. This practical level of chance is defined as the upper confidence interval of a random classifier's accuracy. Given the number of trials, the resulting accuracy of a BCI experiment must be higher than the practical level of chance. Then the BCI can be said to perform significantly better than chance.

The original publication assumes that classes are balanced [48]. In this section, we describe a more general approach that can handle arbitrary class distributions.

17.3.1 Theoretical Level of Random Classification

In order to test classification results for randomness, a sound definition of random classification is required: A random classifier's output is statistically independent from the true class labels.[1] More formally,

$$P(c_e = c \mid c_t) = P(c_e = c), \tag{17.9}$$

where c_e is the estimated class label and c_t is the true class label.

[1] Such randomness is not necessarily caused by the classifier alone. The BCI user failing at the task, electrode failures or inadequate features may all decrease the degree of agreement between the estimated and true class labels. The actual source of randomness is not relevant for this analysis.

The probability of such a random classifier correctly classifying a trial is

$$p_0 = \sum_{c \in C} P(c_e = c) \cdot P(c_t = c), \tag{17.10}$$

where C is the set of all available class labels.

While the probability $P(c_t = c)$ of a trial belonging to class c is determined by the experimental setup, the probability $P(c_e = c)$ of the classifier returning class c needs to be carefully considered. The most conservative approach is to find the highest possible p_0 for a given experiment. This is the case for a classifier that always returns the class that occurs most often. Intuitively, such a classifier would not be considered random since its output is purely deterministic, but the output is independent from the true class labels, thus (17.9) applies.

Alternatively, the values for $P(c_e = c)$ can be calculated from the experimental results using the confusion matrix (17.3). This yields the same p_0 that is used for the calculation of Cohen's κ, which is the theoretical chance level of an actual classifier. This approach is less conservative as the chance level no longer depends on the experimental setup alone, but also on the probability of each class to be selected by the classifier. However, this approach can only be applied after classification has been performed.

17.3.2 Confidence Intervals

Can a BCI identify the user's intended message or command more accurately than chance? This question can be formally defined with a statistical test, in which the null hypothesis H_0 represents the hypothesis that the BCI's classification is not more accurate than a random classifier. As discussed later, performing above chance is a necessary, but not sufficient, condition for an effective BCI. BCIs typically must perform well above chance to be useful. For example, a speller that identifies one of 36 targets with 50 % accuracy would perform much better than chance, but would not allow useful communication. Formally, the hypothesis test can be written as

$$\begin{aligned} H_0 &: p \leq p_0 \\ H_1 &: p > p_0, \end{aligned}$$

where p is the true classification accuracy, and p_0 is the classification accuracy of a random classifier. We compare the one-sided confidence interval of p against the theoretical level of chance, p_0. If p_0 lies outside the confidence interval of p, we can reject H_0 in favor of H_1, thereby indicating that the classifier performs significantly better than chance, at the chosen level of significance.

Regardless of the number of classes, classification can be reduced to either of two outcomes: correct or wrong classification. The correct classification of a trial is called "success." When the probability of success is p, then the probability of getting exactly K successes from N independent trials follows the binomial distribution:

$$f(K;N,p) = \binom{N}{K} p^K (1-p)^{N-K}. \tag{17.11}$$

In a BCI experiment, the classification accuracy is an estimate of p, the true probability of correctly classifying a trial. Given the observed classification accuracy $\hat{p}$, a confidence interval can be calculated that contains the true p with a probability of $1-\alpha$.

Different confidence intervals have been proposed in the literature. The Clopper–Pearson "exact" interval, as well as the Wald interval are too conservative and should not be used in favor of the adjusted Wald interval or the Wilson score interval [8]. We will focus on the adjusted Wald interval because of its simplicity.

17.3.2.1 Adjusted Wald Confidence Interval for Classification Accuracy

Consider the situation where we have N independent trials, of which K are correctly classified. Adding two successes and two failures to the experimental result leads to an unbiased estimator for the probability $\hat{p}$ of correct classification (17.12). Upper and lower confidence limits of $\hat{p}$ are given by (17.13) and (17.14) respectively.

$$\hat{p} = \frac{K+2}{N+4} \tag{17.12}$$

$$p_u = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N+4}} \tag{17.13}$$

$$p_l = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N+4}} \tag{17.14}$$

$z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. For a one-sided confidence interval $z_{1-\alpha}$ can be used instead of $z_{1-\alpha/2}$.

17.3.2.2 Adjusted Wald Confidence Interval for Cohen's Kappa

Kappa is calculated by transforming accuracy values from the interval $[p_0, 1]$ to the interval $[0, 1]$ according to (17.2). Similarly, a confidence interval of the classification accuracy can be transformed, resulting in a confidence interval for κ

$$\kappa_{l/u} = \frac{p_{l/u} - p_0}{1-p_0}. \tag{17.15}$$

This results in a modified null hypothesis that tests κ and associated confidence intervals against zero

$$H_0 : \kappa \leq 0$$
$$H_1 : \kappa > 0.$$

The original publication introducing κ [15] proposes a confidence interval that is derived from the Wald interval, which is too conservative according to [8]. Applying the adjusted Wald interval instead results in (17.16)–(17.19)

$$\hat{p} = \frac{K+2}{N+4} \tag{17.16}$$

$$\hat{\kappa} = \frac{\hat{p} - p_0}{1 - p_0} \tag{17.17}$$

$$\kappa_l = \hat{\kappa} - z_{1-\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{(N+4)(1-p_0)} \tag{17.18}$$

$$\kappa_u = \hat{\kappa} + z_{1-\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{(N+4)(1-p_0)}, \tag{17.19}$$

where $\hat{\kappa}$ is the value of kappa that follows from the unbiased estimator in (17.16).

17.3.3 Summary

It is important not only to consider point estimators of performance measures but also to use appropriate statistics to validate experimental results. In this section we showed how to test estimates of classification accuracy and Cohen's κ against results expected from random classification.

Care has to be taken to chose an appropriate model of random classification. Without knowledge of the classifier's behavior conservative assumptions have to be made about chance classification. When classification results are available a less conservative chance level can be estimated from the classifier output.

17.4 Performance Metrics Incorporating Time

Another critical factor in any communication system is speed—the time required to accomplish a goal, such as spelling a sentence or navigating a room. BCIs often report performance in terms of ITR or bit rate, a common metric for measuring the information sent within a given time [58, 66]. We will measure ITR in bits per minute, and bit rate in bits per trial, which can be calculated via

$$B = \log_2 C + \hat{p} \log_2 \hat{p} + (1 - \hat{p}) \log_2 \frac{1 - \hat{p}}{C - 1}, \tag{17.20}$$

where $\hat{p}$ is the estimated classification accuracy and C is the total number of classes (i.e. possible selections). This equation provides the amount of information (in bits) communicated with a single selection. Many BCI articles multiply B by the number of selections per unit time to attain the ITR, measured in bits per minute. In a trial based BCI this is accomplished by multiplying the ITR by the actual number of trials performed per minute.

However, in a typical BCI speller, users correct errors through a "backspace" function, which may be activated manually or automatically via detection of a neuronal error potential [56]. In contrast to the ITR, the WSR (17.21)–(17.22) incorporates such error correction functionality [23].

$$\mathrm{SR} = \frac{B}{\log_2 C} \tag{17.21}$$

$$\mathrm{WSR} = \begin{cases} (2\mathrm{SR} - 1)/T & \mathrm{SR} > 0.5 \\ 0 & \mathrm{SR} \leq 0.5 \end{cases}, \tag{17.22}$$

where SR is referred to as symbol rate, and T is the trial duration in minutes (including eventual delays).

The WSR incorporates correction of an error by two additional selections (backspace and new selection). However, another error may happen during the correction process. This has been addressed by the practical bit rate (PBR) [61], calculated via

$$\mathrm{PBR} = \begin{cases} B(2p - 1)/T & \hat{p} > 0.5 \\ 0 & \hat{p} \leq 0.5 \end{cases}. \tag{17.23}$$

However, WSR or PBR may not be suitable for systems that use other mechanisms to correct errors [1, 17], or if the user chooses to ignore some or all errors.

ITR calculation may seem to rest on a few simple formulae. However, ITR is often misreported, partly to exaggerate a BCI's performance and partly due to inadequate understanding of many assumptions underlying ITR calculation. Articles that only report the time required to convey a single message or command might ignore many delays that are inevitable in realworld BCI operation. BCIs often entail delays between selections for many reasons. A BCI system might need time to process data to reach a classification decision, present feedback to the user, clear the screen, allow the user to choose a new target, and/or provide a cue that the next trial will begin. Delays also occur if a user decides to correct errors.

Moreover, various factors could affect the effective information transfer rate [2], which incorporates advanced features that could help users attain goals more quickly without improving the raw bit rate. Some BCIs may feature automatic tools to correct errors or complete words or sentences. These tools may introduce some delays, which are presumably welcome because they avoid the greater delays that might be necessary to manually correct errors to complete their messages. Similarly, some BCIs may focus on goal-oriented selections rather than process-oriented selections [1, 64]. Consider two BCIs that allow a user to choose one of eight items with perfect accuracy every ten seconds. Each BCI has a raw ITR of 18 bits/min. However, the first BCI allows a user to move a wheelchair one meter in one of eight directions with each selection, and a second BCI might instead let users choose a target room (leaving the system to work out the details necessary to get there). Other BCIs might incorporate context in various ways. BCIs might change the mapping from brain signals to outcomes. For example, if a robot is in an open space, then imagining left hand movement could move the robot left, but if a wall is to the robot's left, then the same mental command could instruct the robot to follow the wall [42]. BCIs could also use context to change the options available to a user. For example, if a user turns a light off, or if the light bulb burns out, then the option of turning on a light might simply not be available [1].

Moreover, ITR has other limitations [4]. For example, ITR is only meaningful for some types of BCIs. ITR is best suited to synchronous BCIs. Self-paced BCIs, in which the user can freely choose when to make selections or refrain from communicating, are not well suited to ITR estimation. ITR also does not account for different types of errors, such as false positives vs. misses, which could influence the time needed for error correction. Reporting ITR might encourage developers to focus on maximizing ITR, even though some users may prefer higher accuracy, even if it reduces ITR.

In summary, ITR calculation is more complicated than it may seem. Articles that report ITR should include realworld delays, account for tools that might increase effective ITR, and consider whether ITR is the best metric. In some cases, articles present different ITR calculation methods such as practical bit rate or raw bit rate [33, 63]. In such cases, authors should clearly specify the differences in ITR calculation methods and explain why different methods were explored.

17.5 Estimating Performance Measures on Offline Data

BCI researchers often perform initial analysis on offline data to test out a new approach, e.g. a new signal processing method, a new control paradigm etc. For example, [4, 10] report on offline results of a hybrid feature set before they apply it in an online BCI [11].

Because the data is available offline it may be manipulated in a way that is not possible with online data. Common manipulations used in the analysis of offline data include, but are not limited to, cross validation, iteration over a parameter space and the use of machine learning techniques. When applying any offline analysis method, it is important to consider firstly the statistical significance of the reported results and secondly how well the results translate to online BCI operation.

Statistical significance must be reported on the results of classifying a dataset which is separate from the dataset used to train the classification function. The dataset the results are reported on is referred to as the verification (or testing) set while the dataset the classifier is trained on is referred to as the training set. Separating training and verification sets allows us to estimate the expected performance of the trained classifier on unseen data.

The ability to translate offline analysis results to online BCI operation depends on a number of factors including the effects of feedback in online analysis, any temporal drift effects in the signal and how well the offline analysis method is constructed to ensure that the results generalize well. These issues will be considered further in the subsequent sections.

17.5.1 Dataset Manipulations

In online BCI operation any parameters (e.g. classifier weights, feature indices) must be learned first before operation of the BCI begins. However with offline data the trials within the dataset may be manipulated freely.

The most straightforward approach is to simply split the dataset into a training and validation set. This could be done with or without re-sorting the trials. If no re-sorting is used and the trials at the beginning of the session are used for training, this is analogous to online analysis. On the other hand it may be desirable to remove serial regularities from the dataset via re-sorting the trial order prior to splitting into training and validation sets.

A common approach taken is to use either k-fold or leave 1 out cross validation. In k-fold cross validation the dataset is split into K subsets. One of these subsets (subset l) is omitted (this is denoted as the "hold out" set), the remainder are used to train the classifier function. The trained function is then used to classify trials in the lth hold out set. This operation is repeated K times with each set being omitted once. Leave 1 out cross validation is identical, except that each hold out set contains just one trial. Thus, every trial is omitted from the training set once.

Cross validation requires trials to be independent. In general this is not the case due to slowly varying baseline, background activity and noise influence. Trials recorded close to each other are likely to be more similar than trials recorded further apart in time. This issue is addressed through h-block cross validation [39]. h trials closest to each trial in the validation set are left out of the training set, in order to avoid overfitting due to temporal similarities in trials.

Another approach taken, particularly in situations where the size of the available dataset is small, is to use bootstrapping. The training (and possibly the validation) set is created from bootstrap replications of the original dataset. A bootstrap replication is a new trial created from the original dataset in such a way that it preserves some statistical or morphological properties of the original trial. For example, [40] describes a method to increase the training set for BCI operation by randomly swapping features between a small number of original trials to create a much larger set of bootstrap replications.

17.5.2 Considerations

Ultimately, the results reported from offline analysis should readily translate to online BCI operation. Therefore when deciding on any data manipulations and/or machine learning techniques the following considerations should be made:

1. Temporal drift in the dataset. During online BCI operation, factors such as fatigue, learning and motivation affect the ability of the BCI user to exert control. If trials are randomly re-sorted in offline analysis the effect of such temporally dependent changes in the signal are destroyed.
2. The effects of feedback. During online BCI operation the classifier results are fed back to the user via exerted control. This affects the users' motivation and hence the signals recorded from them.
3. Overlearning and stability. Classification methods should be stable when applied to large datasets recorded over prolonged periods of time. Thus, efforts must be made to ensure manipulations made to datasets during offline analysis do not lead to an overlearning effect resulting in poor generalisation and performance instability.

17.6 Hypothesis Testing

Statistical significance of the results obtained in a study is reported via testing against a null hypothesis (H_0), which corresponds to a general or default position (generally the opposite of an expected or desired outcome). For example, in studies reporting classification accuracies, the null hypothesis is that classification is random, i. e. the classification result is uncorrelated with the class labels (see Sect. 17.3). In Sect. 17.3 we discussed the use of confidence intervals in testing against the null hypothesis. This section will elaborate further on additional approaches to testing against the null hypothesis, issues that may arise, and how to properly report results.

Many BCI papers present new or improved methods such as new signal processing methods, new pattern recognition methods, or new paradigms, aiming to improve overall BCI performance. From a scientific point of view, the statement that one method is better than another method is only justifiable if it is based on a solid statistical analysis.

A prerequisite for all statistical tests described in the following subsections is a sufficiently large sample size. The optimal sample size depends on the level of the

α (type I) and β (type II) error and the effect size ϵ [9]. The effect size refers to the magnitude of the smallest effect in the population which is still of substantive significance (given the alternate hypothesis H_1 is valid). For smaller effect sizes, bigger sample sizes are needed and vice versa. Cohen suggested values for small, medium, and large effect sizes and their corresponding sample sizes [16].

The following guidelines are a rough summary of commonly used statistical tests for comparing different methods and should help in finding an appropriate method for the statistical analysis of BCI performance.

17.6.1 Student's t-Test vs. ANOVA

To find out if there is a significant difference in performance between two methods, a Student's t-test is the statistic of choice. However, this does not apply to the case where more than two methods should be compared. The reason for this is that every statistical test has a certain probability of producing an error of Type I—that is, incorrectly rejecting the null hypothesis. In the case of the t-test, this would mean that the test indicates a significant difference, although there is no difference in the population (this is referred to as the type I error, false positives, or α error). For t-test we establish an upper bound on the probability of producing an error of Type I. This is the significance level of the test, denoted by the p-value. For instance, a test with $p \leq 0.04$ indicates the probability of a Type I error is no greater than 4 %. If more than one t-test is calculated, this Type I (α) error probability accumulates over independent tests.

There are two ways to cope with this α-error accumulation. Firstly, a correction for multiple testing such as Bonferroni correction could be applied (see Sect. 17.6.3). Secondly, an analysis of variances (ANOVA) with an adequate post-hoc test avoids the problem of α-error accumulation. The advantage of an ANOVA is that it does not perform multiple tests, and in case of more than one factor or independent variable interactions between these variables can also be revealed (see Sect. 17.6.2).

17.6.2 Repeated Measures

There are different ways to study the effects of new methods. One way is to compare the methods by applying each method to a separate subgroup of one sample, meaning every participant is only tested with one method. Another way is to apply each method to every participant, meaning that each participant is tested repeatedly. For statistical analysis, the way the data has been collected must be considered. In case of repeated measures, different statistical tests must be used as compared to separated subgroups. For a regular t-test and an ANOVA, it is assumed that the samples are independent, which is not fulfilled if the same participants are measured repeatedly. A t-test for dependent samples and an ANOVA for repeated measures take the dependency of the subgroups into account and are therefore the methods of choice for repeated measurements.

In a repeated measures ANOVA design, the data must satisfy the sphericity assumption. This has to be verified (i.e. Mauchly's test for sphericity), and if the assumption is violated, a correction method such as Greenhouse Geisser correction must be applied. Most statistical software packages provide tests for sphericity and possible corrections.

In summary, comparing different methods on the same data set also requires repeated measures tests, which is the classical setting for most offline analyses.

17.6.3 Multiple Comparisons

Section 17.3.2 showed how to test a single classification result against the null hypothesis of random classification. This approach is adequate when reporting a single classification accuracy. However, consider the case when multiple classifications are attempted simultaneously. For example, if one has a dataset containing feature sets spanning a range of different time-frequency locations, one may train a classifier on each feature independently and report significantly better then chance performance, at a desired significance level (e.g. $p < 0.05$), if at least one of these classifiers perform better then chance. In this case, the probability of us falsely reporting better then chance performance for a single classifier is 5 % (the Type 1 error rate). However, if we have 100 classifiers each being trained on an independent feature, then we would expect on average five (5 %) of these classifiers to falsely appear to perform significantly better than chance. Thus, if fewer than six of our classifiers independently perform better than chance, we cannot reject the null hypothesis of random classification at the 5 % significance level.

To adjust for this multiple comparisons problem, Bonferroni correction is commonly applied. This is an attempt to determine the family-wise error rate (the probability of making a type 1 error when multiple hypotheses are tested simultaneously). For n comparisons, the significance level is adjusted by $1/n$. Thus, if 100 independent statistical tests are carried out simultaneously, the significance level for each test is multiplied by 1/100. In our previous example, our original significance level of 0.05 (5 %) would thus be reduced to $0.05/100 = 0.0005$. If we were then to select any single classifier which performs significantly better than chance at this adjusted significance level, we may be confident that in practical application it could be expected to perform better than chance at the 5 % significance level.

BCI studies often report features identified in biosignals which may be useful for BCI control. These signals produce very high-dimensional feature spaces due to the combinatorial explosion of temporal, spatial or spectral dimensions. Traditional analysis methods suggest that it is necessary to correct for multiple comparisons.

However, often in biomedical signal processing such corrections prove to be too conservative.

For example, in a plethora of studies from multiple labs, features derived from the event related desynchronization (ERD) have been successfully shown to reliably allow control of BCIs via imagined movement (see for example [20,35,41,50,54]).

However, if one attempts to report the statistical significance of the ERD effect in the time-frequency spectra—treating every time-frequency location as an independent univariate test—using Bonferroni correction, the effect may not pass the test of statistical significance. Say, for example, we observe an ERD effect in a set of time-frequency features spanning a 2 s interval (sampled as 250 Hz) and a frequency range of 1–40 Hz, in 1 Hz increments. Say also we have 100 trials, 50 of which contain the ERD effect and 50 of which do not. Our dataset contains $250 \times 2 \times 40$ features and we are interested in which of them contain a statistically significant difference between the 50 trials in which an ERD is observed and the 50 trials in which an ERD is not observed. We are making 20,000 comparisons, therefore the Bonferroni adjustment to our significance level is 1/20,000. With such a large adjustment, we find that classifiers trained on those time-frequency features encompassing the ERD do not exhibit performance surpassing this stringent threshold for significance. In fact, with this many comparisons, if we wished to continue using Bonferroni correction, we would need a much larger number of trials before we began to see a significant effect.

This highlights a fundamental issue with applying Bonferroni correction to BCI features. Namely, the Bonferroni correction assumes independence of the comparisons. This is an adequate assumption when considering coin tosses (and a number of other more interesting experimental paradigms). However, the biosignals used for BCI classification features, typically derived from co-dependent temporal, spatial, and spectral dimensions of the signal, cannot be assumed to be independent. This must be taken into account when correcting for multiple comparisons.

The false discovery rate (FDR) has been proposed to allow multiple comparison control that is less conservative than the Bonferroni correction, particularly in cases where the individual tests may not be independent. This comes at the risk of increased likelihood of Type 1 errors. The proportion of false positives is controlled instead of the probability of a single false positive. This approach is routinely used to control for Type I errors in functional magnetic resonance tomography (fMRI) maps, EEG/MEG, and functional near infrared spectroscopy (fNIRS) (see for example [14, 26, 28]). However, dependencies between time, frequency and spatial locations may not be adequately accounted for.

A new hierarchical significance testing approach proposed in [59] may provide a solution. The EEG is broken into a time-frequency hierarchy. For example, a family of EEG features at different time-frequency locations may be broken into frequency band sub-families (child hypothesis). Each of these frequency families may be further deconstructed into time sub-families. Hypothesis testing proceeds down the tree with pruning at each node of the tree if we fail to reject the null hypothesis at that node. Child hypotheses are recursively checked if their parents' null hypothesis is rejected. This pruning approach prevents the multiple comparisons correction from being overly conservative while accounting for time-frequency dependencies.

17.6.4 Reporting Results

To correctly report the results of a statistical analysis, the values of the test statistic (t-test: t-value, ANOVA: F-value), the degrees of freedom (subscripted to values of test statistic, e. g. t_{df}, $F_{\mathrm{df1,df2}}$, where df1 stands for in between degrees of freedom and df2 equals within degrees of freedom), and the significance level p (e.g. $p = 0.0008$, $p < 0.05$; $p < 0.01$; $p < 0.001$; or n. s. for not significant results) must be provided. If tests for the violation of assumptions (such as sphericity or normality) are applied, results of these tests and adequate corrections should be reported too.

17.7 Conclusion

A BCI is applied for online control of a computer or device. Yet, offline analysis, including preliminary analyses and parameter optimization, remains an important tool in successful development of online BCI technology. Special care must be taken so that offline analysis readily translates to accurate online BCI operation. Effects from temporal drift in the data, feedback which may not be available in training data, and the possibility of overfitting have to be considered.

A number of different metrics for reporting classification performance are available. From these, classification accuracy is probably the most comprehensible, as it directly corresponds to the probability of performing a correct classification. However, reporting only the accuracy is not sufficient. Depending on the number and distribution of classes, even bad performance can lead to high accuracy values. Therefore, the theoretical chance level and confidence interval should always be reported along with accuracy metrics. Additionally, confusion matrices or ROC curves may provide a more complete picture of classification performance.

When reporting performance metrics that incorporate time, one should always take into account the actual time required to reach a certain goal. This includes trial duration, repetitions, error correction, delays in processing or feedback, and even breaks between trials. Furthermore, this time may be reduced by application specific tools. For instance, consider a BCI spelling system. The time required to spell a complete sentence is likely to be the most important criteria for the BCI user. The bit rate measures the amount of information provided by a single trial, and bit rate multiplied by the rate at which trials are repeated allows one to determine the speed at which individual letters can be spelled. Finally, automatic word completion may reduce the time required to complete words and sentences.

Ultimately, as in almost every other applied science, results of a BCI study will need to be subject to a statistical test. Researchers often seek to demonstrate that a BCI can operate at a particular performance level. Or to demonstrate improved performance of a new method over a previously published method, or compare BCI performance in one population to that of a control group. An appropriate statistic, such as a t-test or ANOVA with or without repeated measures design must be chosen, and when necessary, care should be taken to account for multiple comparisons.

Acknowledgements The views and the conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the corresponding funding agencies.

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement 248320. In addition, the authors would like to acknowledge the following projects and funding sources:

- Coupling Measures for BCIs, FWF project P 20848-N15
- TOBI: Tools for Brain–Computer Interaction, EU project D-1709000020
- Grant National Natural Science Foundation of China, grant no. 61074113.

We would like to express our gratitude towards the reviewers, who provided invaluable thorough and constructive feedback to improve the quality of this chapter.

References

1. Allison, B.Z.: The I of BCIs: Next Generation Interfaces for Brain-Computer Interface Systems That Adapt to Individual Users. Human-Computer Interaction. Novel Interaction Methods and Techniques, vol. 5611, pp. 558–568. Springer Berlin/Heidelberg (2009)
2. Allison, B.Z.: Toward Ubiquitous BCIs. Brain-Computer Interfaces. The Frontiers Collection, pp. 357–387. Springer Berlin/Heidelberg (2010)
3. Allison, B.Z., Neuper, C.: Could Anyone Use a BCI? Brain-Computer Interfaces. Human-Computer Interaction Series, pp. 35–54. Springer London (2010)
4. Allison, B.Z., Brunner, C., Kaiser, V., Müller-Putz, G.R., Neuper, C., Pfurtscheller, G.: Toward a hybrid brain–computer interface based on imagined movement and visual attention. J. Neural Eng. **7**, 026,007 (2010). DOI 10.1088/1741-2560/7/2/026007
5. Atum, Y., Gareis, I., Gentiletti, G., Ruben, A., Rufiner, L.: Genetic feature selection to optimally detect P300 in brain computer interfaces. In: 32nd Annual International Conference of the IEEE EMBS (2010)
6. Bin, G., Gao, X., Wang, Y., Li, Y., Hong, B., Gao, S.: A high-speed BCI based on code modulation VEP. J. Neural Eng. **8**, 025,015 (2011). DOI 10.1088/1741-2560/8/2/025015
7. Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., Flor, H.: A spelling device for the paralysed. Nature **398**, 297–298 (1999). DOI 10.1038/18581
8. Boomsma, A.: Confidence intervals for a binomial proportion. Unpublished manuscript, university of Groningen, Department of Statistics & Measurement Theory (2005)
9. Bortz, J.: Statistik für Sozialwissenschaftler. Springer, Berlin, Heidelberg, New York (1999)
10. Brunner, C., Allison, B.Z., Krusienski, D.J., Kaiser, V., Müller-Putz, G.R., Pfurtscheller, G., Neuper, C.: Improved signal processing approaches in an offline simulation of a hybrid brain–computer interface. J. Neurosci. Methods **188**, 165–173 (2010). DOI 10.1016/j.jneumeth.2010.02.002
11. Brunner, C., Allison, B.Z., Altstätter, C., Neuper, C.: A comparison of three brain–computer interfaces based on event-related desynchronization, steady state visual evoked potentials,

or a hybrid approach using both signals. J. Neural Eng. **8**, 025,010 (2011a). DOI 10.1088/1741-2560/8/2/025010

12. Brunner, P., Ritaccio, A.L., Emrich, J.F., Bischof, H., Schalk, G.: Rapid communication with a "P300" matrix speller using electrocorticographic signals (ECoG). Front. Neurosci. **5**, 5 (2011b)
13. Cabestaing, F., Vaughan, T.M., Mcfarland, D.J., Wolpaw, J.R.: Classification of evoked potentials by Pearson's correlation in a brain–computer interface. Matrix **67**, 156–166 (2007)
14. Chumbley, J.R., Friston, K.J.: False discovery rate revisited: FDR and topological inference using gaussian random fields. NeuroImage **44**(1), 62–70 (2009). DOI 10.1016/j.neuroimage.2008.05.021, http://www.ncbi.nlm.nih.gov/pubmed/18603449
15. Cohen, J.: A coefficient of agreement for nominal scales. Psychol. Meas. **20**, 37–46 (1960)
16. Cohen, J.: A power primer. Psychol. Bull. **112**(1), 155–159 (1992)
17. Dal Seno, B. Matteucci, M., Mainardi, L.: Online detection of P300 and error potentials in a BCI speller. Computational Intelligence and Neuroscience, pp. 1–5 (2010)
18. Daly, I., Nasuto, S., Warwick, K.: Single tap identification for fast BCI control. Cogn. Neurodyn. **5**, 21–30 (2011)
19. Dornhege, G., del R Millán, J., Hinterberger, T., McFarland, D.J., Müller, K.R.: (eds.) Towards Brain–Computer Interfacing. MIT Press (2007)
20. Eskandari, P., Erfanian, A.: Improving the performance of brain–computer interface through meditation practicing. In: Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, pp. 662–665 (2008). DOI 10.1109/IEMBS.2008.4649239
21. Falk, T., Paton, K., Power, S., Chau, T.: Improving the performance of NIRS-based brain–computer interfaces in the presence of background auditory distractions. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pp. 517–520 (2010). DOI 10.1109/ICASSP.2010.5495643
22. Farwell, L.A., Donchin, E.: Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalogr. Clin. Neurophysiol. **70**, 510–523 (1988)
23. Furdea, A., Halder, S., Krusienski, D.J., Bross, D., Nijboer, F., Birbaumer, N., Kübler, A.: An auditory oddball (P300) spelling system for brain–computer interfaces. Psychophysiology **46**, 1–9 (2009). DOI 10.1111/j.1469-8986.2008.00783.x
24. Galán, F., Nuttin, M., Lew, E., Ferrez, P.W., Vanacker, G., Philips, J., del R Millán, J.: A brain-actuated wheelchair: asynchronous and non-invasive brain–computer interfaces for continuous control of robots. Clin. Neurophysiol. **119**, 2159–2169 (2008). DOI 10.1016/j.clinph.2008.06.001
25. Gareis, I., Gentiletti, G., Acevedo, R., Rufiner, L.: Feature extraction on brain computer interfaces using discrete dyadic wavelet transform: preliminary results. Journal of Physics: Conference Series (IOP) **313**, pp. 1–7 (2011)
26. Genovese, C., Wasserman, L.: Operating characteristics and extensions of the false discovery rate procedure. J. R. Stat. Soc. Series B Stat. Methodol. **64**(3), 499–517 (2002). DOI 10.1111/1467-9868.00347, http://doi.wiley.com/10.1111/1467-9868.00347
27. Guger, C., Ramoser, H., Pfurtscheller, G.: Real-time EEG analysis with subject-specific spatial patterns for a brain–computer interface (BCI). IEEE Trans. Neural Syst. Rehabil. Eng. **8**, 447–450 (2000). DOI 10.1109/86.895947
28. Hemmelmann, C., Horn, M., Süsse, T., Vollandt, R., Weiss, S.: New concepts of multiple tests and their use for evaluating high-dimensional EEG data. J. Neurosci. Methods **142**(2), 209–17 (2005). DOI 10.1016/j.jneumeth.2004.08.008, http://ukpmc.ac.uk/abstract/MED/15698661/reload=1
29. Hild II, K.E., Kurimo, M., Calhoun, V.D.: The sixth annual MLSP competition, 2010. Machine Learning for Signal Proc (MLSP '10) (2010)
30. Hoffmann, U., Vesin, J.M., Ebrahimi, T., Diserens, K.: An efficient P300-based brain–computer interface for disabled subjects. J. Neurosci. Methods **167**, 115–125 (2008). DOI 10.1016/j.jneumeth.2007.03.005

31. Horki, P., Solis-Escalante, T., Neuper, C., Müller-Putz, G.: Combined motor imagery and SSVEP based BCI control of a 2 DoF artificial upper limb. Med. Biol. Eng. Comput. (2011). DOI 10.1007/s11517-011-0750-2
32. Huggins, J.E., Levine, S.P., BeMent, S.L., Kushwaha, R.K., Schuh, L.A., Passaro, E.A., Rohde, M.M., Ross, D.A., Elisevich, K.V., Smith, B.J.: Detection of event-related potentials for development of a direct brain interface. J. Clin. Neurophysiol. **16**(5), 448 (1999)
33. Jin, J., Allison, B., Sellers, E., Brunner, C., Horki, P., Wang, X., Neuper, C.: Optimized stimulus presentation patterns for an event-related potential EEG-based brain–computer interface. Med. Biol. Eng. Comput. **49**, 181–191 (2011). doi:10.1007/s11517-010-0689-8
34. Kalcher, J., Flotzinger, D., Neuper, C., Gölly, S., Pfurtscheller, G.: Graz brain–computer interface II: towards communication between humans and computers based on online classification of three different EEG patterns. Med. Biol. Eng. Comput. **34**, 382–388 (1996). DOI 10.1007/BF02520010
35. Karrasch, M., Laine, M., Rapinoja, P., Krause, C.M.: Effects of normal aging on event-related desynchronization/synchronization during a memory task in humans. Neurosci. Lett. **366**(1), 18–23 (2004). DOI 10.1016/j.neulet.2004.05.010, http://dx.doi.org/10.1016/j.neulet.2004.05.010
36. Krausz, G., Ortner, R., Opisso, E.: Accuracy of a brain computer interface (p300 spelling device) used by people with motor impairments. Stud. Health Technol. Inform. **167**, 182–186 (2011)
37. Kübler, A., Birbaumer, N.: Brain-computer interfaces and communication in paralysis: extinction of goal directed thinking in completely paralysed patients? Clin. Neurophysiol. **119**, 2658–2666 (2008). DOI 10.1016/j.clinph.2008.06.019
38. Kübler, A., Nijboer, F., Mellinger, J., Vaughan, T.M., Pawelzik, H., Schalk, G., McFarland, D.J., Birbaumer, N., Wolpaw, J.R.: Patients with ALS can use sensorimotor rhythms to operate a braincomputer interface. Neurology **64**, 1775–1777 (2005)
39. Lemm, S., Blankertz, B. Dickhaus, T., Müller, K.R.: Introduction to machine learning for brain imaging. NeuroImage **56**(2), pp. 387–399 (2011)
40. Lotte, F.: Generating artificial EEG signals to reduce BCI calibration time. In: Proceedings of the 5th International Brain–Computer Interface Conference 2011, pp. 176–179 (2011)
41. Mason, S.G., Birch, G.E.: A brain-controlled switch for asynchronous control applications. IEEE Trans. Biomed. Eng. **47**, 1297–1307 (2000)
42. Millán, J., Mouriño, J.: Asynchronous BCI and local neural classifiers: an overview of the adaptive brain interface project. IEEE Trans. Neural Syst. Rehabil. Eng. **11**, 159–161 (2003)
43. Millán, J., Mouriño, J., Franzé M., Cincotti, F., Varsta, M., Heikkonen, J., Babiloni, F.: A local neural classifier for the recognition of EEG patterns associated to mental tasks. IEEE Trans. Neural Netw. **13**, 678–686 (2002)
44. Müller, K.R., Anderson, C.W., Birch, G.E.: Linear and nonlinear methods for brain–computer interfaces. IEEE Trans. Neural Syst. Rehabil. Eng. **11**, 165–169 (2003)
45. Müller, K.R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., Blankertz, B.: Machine learning for real-time single-trial EEG analysis: from brain–computer interfacing to mental state monitoring. J. Neurosci. Meth. **167**, 82–90 (2008). DOI 10.1016/j.jneumeth.2007.09.022
46. Müller-Putz, G.R., Pfurtscheller, G.: Control of an electrical prosthesis with an SSVEP-based BCI. IEEE Trans. Biomed. Eng. **55**, 361–364 (2008). DOI 10.1109/TBME.2007.897815
47. Müller-Putz, G.R., Scherer, R., Pfurtscheller, G., Rupp, R.: EEG-based neuroprosthesis control: a step towards clinical practice. Neurosci. Lett. **382**, 169–174 (2005)
48. Müller-Putz, G.R., Scherer, R., Brunner, C., Leeb, R., Pfurtscheller, G.: Better than random? A closer look on BCI results. Int. J. Bioelectromagn. **10**, 52–55 (2008)
49. Neuper, C., Müller, G.R., Kübler, A., Birbaumer, N., Pfurtscheller, G.: Clinical application of an EEG-based brain–computer interface: a case study in a patient with severe motor impairment. Clin. Neurophysiol. **114**, 399–409 (2003)
50. Pfurtscheller, G., Neuper, C.: Motor imagery and direct brain–computer communication. Proc. IEEE **89**, 1123–1134 (2001). DOI 10.1109/5.939829

51. Pfurtscheller, G., Müller, G.R., Pfurtscheller, J., Gerner, H.J., Rupp, R.: "Thought"-control of functional electrical stimulation to restore handgrasp in a patient with tetraplegia. Neurosci. Lett. **351**, 33–36 (2003). DOI 10.1016/S0304-3940(03)00947-9
52. Piccione, F., Giorgi, F., Tonin, P., Priftis, K., Giove, S., Silvoni, S., Palmas, G., Beverina, F.: P300-based brain computer interface: reliability and performance in healthy and paralysed participants. Clin. Neurophysiol. **117**, 531–537 (2006). DOI 10.1016/j.clinph.2005.07.024
53. Rebsamen, B., Guan, C., Zhang, H., Wang, C., Teo, C., Ang, M.H., Burdet, E.: A brain controlled wheelchair to navigate in familiar environments. IEEE Trans. Neural Syst. Rehabil. Eng. **18**(6), 590–598 (2010). DOI 10.1109/TNSRE.2010.2049862, http://dx.doi.org/10.1109/TNSRE.2010.2049862
54. Roberts, S., Penny, W., Rezek, I.: Temporal and spatial complexity measures for electroencephalogram based brain–computer interfacing. Med. Biol. Eng. Comput. **37**, 93–98 (1999). doi:10.1007/BF02513272
55. Ryan, D.B., Frye, G.E., Townsend, G., Berry, D.R., Mesa-G, S., Gates, N.A., Sellers, E.W.: Predictive spelling with a P300-based brain–computer interface: Increasing the rate of communication. Int. J. Hum. Comput. Interact. **27**, 69–84 (2011). DOI 10.1080/10447318.2011.535754
56. Schalk, G., Wolpaw, J.R., McFarland, D.J., Pfurtscheller, G.: EEG-based communication: presence of an error potential. Clin. Neurophysiol. **111**, 2138–2144 (2000)
57. Schlögl, A., Kronegg, J., Huggins, J.E., Mason, S.G.: Evaluation criteria for BCI research. In: Toward brain–computer interfacing. MIT Press (2007)
58. Shannon, C.E., Weaver, W.: A mathematical theory of communication. University of Illinois Press (1964)
59. Singh, A.K., Phillips, S.: Hierarchical control of false discovery rate for phase locking measures of EEG synchrony. NeuroImage **50**(1), 40–47 (2010). DOI 10.1016/j.neuroimage.2009.12.030, http://dx.doi.org/10.1016/j.neuroimage.2009.12.030
60. Sitaram, R., Zhang, H., Guan, C., Thulasidas, M., Hoshi, Y., Ishikawa, A., Shimizu, K., Birbaumer, N.: Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface. NeuroImage **34**, 1416–1427 (2007)
61. Townsend, G., LaPallo, B.K., Boulay, C.B., Krusienski, D.J., Frye, G.E., Hauser, C.K., Schwartz, N.E., Vaughan, T.M., Wolpaw, J.R., Sellers, E.W.: A novel P300-based brain–computer interface stimulus presentation paradigm: Moving beyond rows and columns. Clin. Neurophysiol. **121**, 1109–1120 (2010)
62. Vidaurre, C., Blankertz, B.: Towards a cure for BCI illiteracy. Brain Topogr. **23**, 194–198 (2010). DOI 10.1007/s10548-009-0121-6
63. Volosyak, I.: SSVEP-based Bremen-BCI interface – boosting information transfer rates. J. Neural Eng. **8**, 036,020 (2011). DOI 10.1088/1741-2560/8/3/036020
64. Wolpaw, J.R.: Brain-computer interfaces as new brain output pathways. J. Physiol. **579**, 623–619 (2007). DOI 10.1113/jphysiol.2006.125948
65. Wolpaw, J.R., Flotzinger, D., Pfurtscheller, G., McFarland, D.J.: Timing of EEG-based cursor control. J. Clin. Neurophysiol. **14**(6), 529–538 (1997)
66. Wolpaw, J.R., Birbaumer, N., Heetderks, W.J., McFarland, D.J., Peckham, P.H., Schalk, G., Donchin, E., Quatrano, L.A., Robinson, C.J., Vaughan, T.M.: Brain-computer interface technology: a review of the first international meeting. IEEE Trans. Rehabil. Eng. **8**, 164–173 (2000). DOI 10.1109/TRE.2000.847807
67. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. Clin. Neurophysiol. **113**, 767–791 (2002). DOI 10.1016/S1388-2457(02)00057-3