# The BCI Competition III:
## Validating Alternative Approaches to Actual BCI Problems

Benjamin Blankertz, Klaus-Robert Müller, Dean Krusienski, Gerwin Schalk, Jonathan R. Wolpaw, Alois Schlögl, Gert Pfurtscheller, José del R. Millán, Michael Schröder, Niels Birbaumer

***Abstract*— A Brain-Computer Interface (BCI) is a system that allows its users to control external devices with brain activity. Although the proof-of-concept was given decades ago, the reliable translation of user intent into device control commands is still a major challenge. Success requires the effective interaction of two adaptive controllers: the user's brain, which produces brain activity that encodes intent, and the BCI system, which translates that activity into device control commands. In order to facilitate this interaction, many laboratories are exploring a variety of signal analysis techniques to improve the adaptation of the BCI system to the user. In the literature, many machine learning and pattern classification algorithms have been reported to give impressive results when applied to BCI data in offline analyses. However, it is more difficult to evaluate their relative value for actual online use. BCI data competitions have been organized to provide objective formal evaluations of alternative methods. Prompted by the great interest in the first two BCI Competitions, we organized the third BCI Competition to address several of the most difficult and important analysis problems in BCI research. This article describes the data sets that were provided to the competitors and gives an overview of the results. In a series of accompanying articles, the winning teams describe their methods in detail.**

***Index Terms*— augmentative communication, beta-rhythm, BCI, brain-computer interface, EEG, ERP, imagined hand movements, mu-rhythm, non-stationarity, P300, rehabilitation, single-trial classification, slow cortical potentials.**

## I. Introduction

BRAIN-COMPUTER INTERFACES (BCIs) allow to directly control a computer application or a technical device by intent alone. The system estimates the intent of the human user from her/his brain signals measured at microscopic, mesoscopic, or macroscopic scale, cf. [1], [2], [3], [4] for an overview. The interest in BCI research is strongly increasing as reflected by the exponentially growing number of published peer-reviewed journal papers on that topic.

BCI Competitions are organized in order to foster the development of improved BCI technology by providing an unbiased validation of a variety of data analysis techniques. In each competition a variety of data sets was made publicly available in a documented format via internet ([5], [6], [7]). Each data set is a record of brain signals from BCI experiments of leading laboratories in BCI technology split into two parts: one part of labeled data ('training set') and another part of unlabeled data ('test set'). Researchers worldwide could tune their methods to the training data and submit the output of their translation algorithms for the test data. The truth about the test data was kept secret until, after the deadline, it was used to evaluate the submissions. This procedure guarantees that the assessment of performance is not biased by overfitting the selection of methods and the choice of their parameters to the data.

The three BCI Competitions were arranged in 2001, 2002 and 2004. The growing interest in such contests is reflected by the number of submissions rising from 10 to 57 to 92. The tasks and results of the first two competitions are summarized in [8], [9]. The first competition was a test for us to see how such an enterprise would work, and how much attention it would attract. In the second competition we provided a broad range of typical fundamental BCI problems. For the third BCI Competition ([7]) presented here we advanced to a diversity of catchy analysis challenges that are highly relevant to present BCI research.

More specifically, the competition comprised the problems of session-to-session transfer, non-stationarity, small training sets, subject-to-subject transfer, continuous test data without trial structure, asynchronous paradigms and idle states.

### A. Ranking of competition results

The ranking of results from Internet competitions cannot be taken at face value since they may not provide a completely objective assessment of quality for several reasons:

(1) There is great variance in how much effort contributors put into preparing their submissions.

(2) When test sets (and the number of classes) are relatively small, luck may also play a big role. For example, if there are 15 methods in a binary problem that are able to classify

Authors BB and KRM were partially supported by grants of the *Bundesministerium für Bildung und Forschung* (BMBF), FKZ 01IBB02A/B and by the *Deutsche Forschungsgemeinschaft* (DFG), FOR 375/B1. Authors DK, GS and JRW's work was supported in part by National Institutes of Health Grants HD30146 (National Center for Medical R ehabilitation Research of the National Institute of Child Health and Human Development) and EB00856 (National Institute of Biomedical Imaging and Bioengineering and National Institute of Neurological Disorders and Stroke ) and the James S. McDonnell Foundation. Author JdRM is supported by the Swiss National Science Foundation NCCR "IM2". Authors BB, KRM and JdRM were partially supported by the PASCAL Network of Excellence, EU # 506778.

BB and KRM are with Fraunhofer FIRST (IDA), Berlin, Germany, E-mail: benjamin.blankertz@first.fraunhofer.de. KRM is also with University of Potsdam, Germany.

DK, GS and JRW are with the Labatory of Nervous System Disorders, Wadsworth Center, New York State Dept. of Health, Albany, NY, USA. JRW is also with the State University of New York, Albany, NY, USA.

AS and GP are with the Institute for Human-Computer Interfaces, University of Technology Graz, Austria.

JdRM is with the IDIAP Research Institute, CH-1920 Martigny, Switzerland

MS is with the Dept. of Technical Computer Science, Eberhard-Karls-Universität Tübingen, Germany.

NB is with the Institute of Medical Psychology and Behavioral Neurobiology, University of Tübingen, Germany and also with the University of Trento, Italy.

TABLE I
IN THIS TABLE THE WINNING TEAMS FOR ALL COMPETITION DATA SETS ARE LISTED. REFER TO SEC. V TO SEE WHY THERE IS NO WINNER FOR DATA SET IVB.

| data set | research lab | contributor(s) |
|---|---|---|
| I | Tsinghua University, Beijing, China | **Qingguo Wei**, Fei Meng, Yijun Wang, Shangkai Gao |
| II | PSI CNRS FRE-2645, INSA de Rouen, France | **Alain Rakotomamonjy**, V. Guigue |
| IIIa | Neural Signal Processing Lab Institute for Infocomm Research, Singapore | **Cuntai Guan**, Haihong Zhang, Yuanqin Li |
| IIIb | Fraunhofer (FIRST) IDA, Berlin, Germany | **Steven Lemm** |
| IVa | Tsinghua University, Beijing, China | **Yijun Wang**, Han Yuan, Dan Zhang, Xiaorong Gao, Zhiguang Zhang, Shangkai Gao |
| IVc | Tsinghua University, Beijing, China | **Dan Zhang**, Yijun Wang |
| V | University of Barcelona | **Ferran Galán**, Francesc Oliva, Joan Guàrdia |

correctly 60 % of the ideal set of all trials with random output on the remaining 40 %, the expected accuracy of all these methods is 80 %. However, on a fixed test set consisting of 100 trials, the expected difference between the best and the worst result is greater than 10 % (assuming independence between methods and test trials).

In Sec. II–VI of this paper, we will describe the eight data sets comprising the competition and we will report and comment on the submissions. The results of all submissions are completely reported on the web ([10]) where we also list short descriptions of all applied methods. A list of the winning teams for each data set is summarized in table I. The winning labs published individual articles on their approaches, see [11], [12], [13], [14], [15], [16], [17] in this issue.

## II. DATA SET I

This data set was provided by the Institute of Medical Psychology and Behavioral Neurobiology, University of Tübingen (head: Niels Birbaumer) and Max-Planck-Institute for Biological Cybernetics, Tübingen, (Bernhard Schökopf), and Universität Bonn, Dept. of Epileptology.

### A. Description of the data set

This data set addresses the robustness of a classification approach. A common task in BCI is to apply a classifier that was trained during previous sessions during a later session without retraining it. The challenge of this task is that the electrical patterns of the patient might show some different characteristics on a new session. This kind of non-stationarity can be caused for example by changed levels of motivation, arousal, fatigue etc. In addition, the recording system might have undergone slight changes concerning electrode positions and impedances.

Data set I reflects this situation: training and test data were recorded from the same subject and the same experimental task, but on two different days with about one week of delay. As electrocorticography (ECoG) was used and not EEG, the variation of electrode positions and impedances are expected to be rather small. The competitors were asked to set up a classifier based on the labeled training data of the first session and apply it to the unlabeled test data of the second session. The performance criteria used for evaluation was the percentage of correctly classified test trials.

The subject was not a locked-In patient but suffered from epilepsy. For this reason his neural activity was monitored for several days with an ECoG recording. During this interval the subject twice participated in a BCI experiment based on motor imagery. The task of both sessions was the same: to produce imagined movements of either the left small finger or the tongue. The provided data sets consist of 278 trials performed during the first session (training data) and 100 trials from the second session (test data). Electrical brain activity was picked up with an $8 \times 8$ ECoG platinum electrode grid which was placed on the contralateral (right) motor cortex. The grid was covered by meninges and scull and was not sensitive to muscle artifacts. As the scull and the meninges act as low-pass filters during EEG recordings, ECoG data can contain stronger high-frequency components than EEG. The grid was assumed to cover the right motor cortex completely, but due to its size (approx. $8 \times 8$ cm) it could in addition record activity from surrounding cortical areas. All recordings were performed with a sampling rate of 1000 Hz. After amplification the recorded potentials were stored as microvolt values.

Trial duration was three seconds. To avoid visually evoked potentials being reflected by the data, the recording intervals started 0.5 seconds after the visual cue had ended. For further information about the experiment, please refer to [18].

### B. Outcome of the competition

We received 27 submissions for the test labels. Many submitted results were of high quality, 12 out of 27 submissions managed to achieve more than 80 percent classification accuracy on the test set. Although including an outlier of only 22 percent accuracy (probably submitted with accidentally confused class labels) the average accuracy of all submissions was 70 percent. Fig. 1 shows the histogram of the submission accuracy.

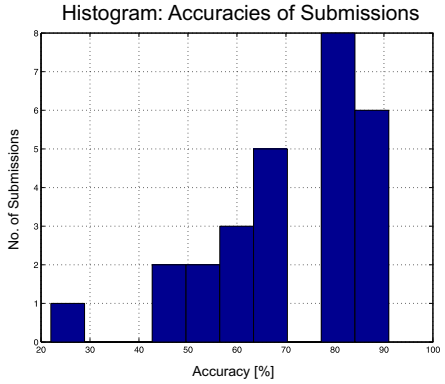


Fig. 1. Histogram of the classification accuracy of 27 submitted solutions. One submission stays clearly below the chance level of 50 percent. A group of 14 submissions reaches more than 78 percent accuracy.

The submissions of rank one to three and their applied methods at a glance:

1) An accuracy of 91 percent was achieved by Qingguo Wei and his co-contributors from the Tsinghua University of Beijing. They used a combination of bandpower features together with CSSD and mean waveforms that were chosen by fisher discriminant analysis before classification was performed with a linear SVM.
2) An accuracy of 87 percent was achieved by Paul Hammon from the University of California in San Diego. After unmixing with ICA, a combination of AR coefficients, spectral power (0-45 Hz) and wavelet coefficients were used as features. Classification was performed with regularized logistic regression.
3) Marginally less, 86 percent accuracy, was reached by three submissions: By Michal Sapinski from Warsaw University, by Mao Dawei and co-contributors from Zhejiang University, and by Alexander D'yakonov from Moscow State University. Their used features comprise the offset and spectral power of hand selected channels (Michal Sapinski), the standard deviation of the Hilbert-Huang Transform for time frequency windows (window size: 5 Hz and 0.2 s) of seven channels (Mao Dawei), and hand chosen features from seven channels (Alexander D'yakonov). For classification, logistic regression (Michal Sapinski) and Mahalanobis distance (Mao Dawei) were used.

Taking a closer look on solutions above 60 percent accuracy, discriminant analysis (linear, robust etc.) dominates the classification methods by 4 entries before (linear) support vector machines with 3 entries. Furthermore logistic regression or mahalanobis/fisher distance was used for two submissions each. Successful methods showed a tendency to use a combination of different feature types.

Fig. 2 takes a closer look onto the difficulty the contributions had with certain test vectors. Most test vectors were classified correctly but around trial 40 (in chronological order, not in competition order), many misclassifications occurred. One interpretation is non-stationarity in the signals caused by eleptiform patterns in the EEG which did arise frequently for this patient.

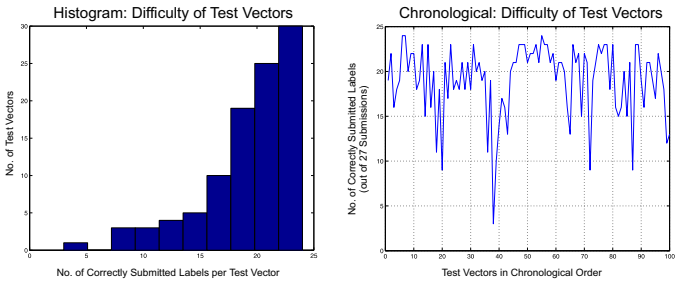


Fig. 2. Difficulty of test vectors from the contributor's point of view. The left histogram shows that no vector was misclassified by every submission and that many vectors received correct labels from 20 or more submissions. Another view on this distribution provides the right graph. It shows the number of correctly submitted labels for every trial in chronological order (the order was randomized for the competition). Around trial no. 40 many trials were not classified correctly.

## III. Data set II: P300 Speller Paradigm

This data set was provided by the Wadsworth Center, New York State Department of Health (head: Jonathan R. Wolpaw).

### A. Description of the data set

This data set represents a complete record of P300 evoked potentials (five sessions from two subjects) recorded with the BCI2000 software [19], using a paradigm described in [20] and originally by Farwell and Donchin [21]. In these experiments, the user was presented with a 6 by 6 matrix of 36 different alphanumeric characters. The user's task was to sequentially focus attention on characters from a word that was defined by the investigator. The 6 rows and 6 columns of this matrix were successively and randomly intensified at a rate of 5.7 Hz. Two out of 12 intensifications of rows or columns highlighted the desired character (i.e., one particular row and one particular column). The responses evoked by these infrequent stimuli (i.e., the 2 out of 12 stimuli that did contain the desired character) are different from those evoked by the stimuli that did not contain the desired character and they are similar to the P300 responses previously reported [20], [21]. Signals from the two subjects were collected from 64 ear-referenced channels (bandpass filtered from 0.1–60 Hz and digitized at 240 Hz) using the BCI2000 software. Each session consisted of nine runs, and each run contained a single word. For each character epoch in the run, the user display was as follows: the matrix was displayed for a 2.5 s period, and during this time each character had the same intensity (i.e., the matrix was blank). Subsequently, each row and column in the matrix was randomly intensified for 100 ms. After intensification of a row/column, the matrix was blank for 75 ms. Row/column intensifications were block randomized in blocks of 12. The sets of 12 intensifications were repeated 15 times for each character epoch (i.e., any specific row/column was intensified 15 times and thus there were 180 total intensifications for each character epoch). Each character epoch was followed by a 2.5 s period, and during this time the matrix was blank. This period informed the user that this character was completed and to focus on the next character in the word that was displayed on the top of the screen (the current character was shown in parentheses). The resulting data for each subject was partitioned into character epochs and divided chronologically into two parts, the first 85 characters for training and the remaining 100 characters for testing. The character epochs in each training and test set were then scrambled to avert identification of the character sequences in the test data. The objective in the contest was to use the 85 characters per subject of training data to construct a classifier, and to then predict the 100 characters per subject in the unlabeled test data. Participants were asked to report the classification results using all 15 flash sequences and, additionally, only the first 5 flash sequences.

### B. Outcome of the competition

A total of 10 submissions were received for this data set, incorporating a wide variety of pre-processing and classification

methods. Using all 15 sequences, the majority of submissions (8) predicted the test characters with at least 75 % accuracy (accuracy expected by chance was 2.8 %). Several contestants achieved an accuracy of over 90 %, and the winner achieved an impressive accuracy of 96.5 % (see [12] for algorithm details).

## IV. Data sets IIIa and IIIb:

This data set is provided by the Institute for Human-Computer Interfaces, University of Technology Graz – BCI Lab (head: Gert Pfurtscheller).

### A. Description of data set IIIa

The data set consists of recordings from 3 subjects; the subjects performed 4 different motor imagery tasks according to a cue. Sixty EEG channels were recorded and the recording was made with a 64-channel EEG amplifier from Neuroscan, using the left mastoid for reference and the right mastoid as ground. The EEG was sampled with 250 Hz, it was filtered between 1 and 50 Hz with Notchfilter on. The data of all runs was concatenated and converted into the GDF format ([22]). The subject sat in a relaxing chair with armrests. The task was to perform imagery left hand, right hand, foot or tongue movements according to a cue. The order of cues was random. The experiment consists of several runs (at least 6) with 40 trials each each; after trial begin, the first 2s were quite, at $t$=2s an acoustic stimulus indicated the beginning of the trial, and a cross '+' is displayed; then from $t$=3s an arrow to the left, right, up or down was displayed for 1s; at the same time the subject was asked to imagine a left hand, right hand, tongue or foot movement, respectively, until the cross disappeared at $t$=7s. Each of the 4 cues was displayed 10 times within each run in a randomized order. Participants should provide a continuous classification output (continuous in time as well as magnitude) for all 4 classes. In other words the classifier should provide 4 continuous traces for the whole data set (including labeled trials, and trials marked as artifact). At each point in time, the trace with the largest value determines the corresponding class. Then, a confusion matrix is built from all trials for each time-point 0.0s $\leq t \leq$ 7.0s . From these confusion matrices, the time course of the accuracy and the time-course of the kappa coefficient can be obtained. The performance measure of the competition was the maximum kappa value in time, averaged for the three subjects.

### B. Outcome of the competition – data set IIIa

We received the following three submissions, whose performance on the competition's test set is shown in table II.

A Authors: Hill & Schröder (Max Planck Institute for Biological Cybernetics, Tübingen and Tübingen University), Method: resampling 100Hz, detrending, Infomax ICA, Amplitude spectra (Welch), linear PCA, and SVM (remark: scores are constant for each trial)

B Authors: Guan, Zhang & Li (Neural Signal Processing Lab Institute for Infocomm Research, Singapore), Method: Fisher ratios of channel-freqency-time bins, feature selection, designing mu- and beta passband, multiclass CSP, SVM

TABLE II
Maximum Kappa for $t \leq$7s in the three subjects (K3, K6, L1) and its mean obtained by the three competitors A, B and C.

| #. | | mean | K3 | K6 | L1 |
|---|---|---|---|---|---|
| 1. | B | 0.79 | 0.82 | 0.76 | 0.80 |
| 2. | C | 0.69 | 0.90 | 0.43 | 0.71 |
| 3. | A | 0.63 | 0.95 | 0.41 | 0.52 |

C Authors: Gao (head), Wu & Wei (Tsinghua University, Beijing, China), Method: surface laplacian, 8-30Hz filter, CSP (one-vs-rest), SVM+kNN+LDA, 'bagging'

A detailed description of the results is available from [23].

### C. Description of data set IIIb

This data set IIIb contained 2-class EEG data from 3 subjects. Each data set contained recordings from consecutive sessions during a BCI experiment. The large amount of data should enable the use of non-stationary classifiers, because it is reasonable to expect that time-varying classifier performs better than a stationary (static) classifier. Moreover, based on the experience of the second BCI competitition [6], [24], [9], the response time of each method has to be evaluated. The experiment consists of 3 sessions for each subject. Each session consists of 4 to 9 runs. The data of all runs was concatenated and converted into the GDF format [22]. The recordings were made with a bipolar EEG amplifier from g.tec. The EEG was sampled with 125 Hz, it was filtered between 0.5 and 30 Hz with Notchfilter on.

In order to evaluate the time delay, it was required that the submitters provided (1) a continuous classification output, and (2) it had to be demonstrated that the used algorithms are causal. The output was validated using the time course of the mutual information [25]. The method with the maximum increase of the mutual information (maximum steepness calculated as $\mathrm{MI}(t)/(t-3s)$ for $t >$3.5s) was used for validation. In order to avoid the involuntary stimulus-response, only time $t >$3.5s was evaluated. The 'steepness' of the mutual information quantifies the response time. The evaluation algorithm is provided in BIOSIG (see `/biosig/t490/criteria2005IIIb.m` in [26]).

### D. Outcome of the competition – data set IIIb

We received seven submissions for this data set. The following three submissions obtained the best performance on the competition's test set, see table III.

A Authors: O.Burmeister, M.Reischl, R. Mikut (Forschungszentrum Karlsruhe, Germany), Method: Bandbower (BP), ratios and differences of BP; MANOVA for feature selection; SVM and linear combiner

C Author: S. Lemm (Fraunhofer-FIRST IDA, Berlin, Germany), Method: ERP and ERD (mu and beta), propabilistic classification model, accumulative classifier

G Authors: Xiaomei Pei, Guangyu Bin (Institute of Biomedical Engineering of Xi'an Jiaotong University, Xi'an,

TABLE III
MAXIMUM STEEPNESS (WITH $t_0$ =3s) OF THE MUTUAL INFORMATION [BITS/S] IN THE THREE SUBJECTS (O3, S4, X11) AND ITS MEAN OBTAINED BY THE THREE COMPETITORS A, B AND C.

| #. | | mean | O3 | S4 | X11 |
|---|---|---|---|---|---|
| 1. | C | 0.32 | 0.17 | 0.44 | 0.35 |
| 2. | A | 0.25 | 0.16 | 0.42 | 0.17 |
| 3. | G | 0.14 | 0.20 | 0.09 | 0.12 |

China), Method: FFT with Hanning window of 1s-segments; Fisher Discriminant Analysis

The main aim was to evaluate causal algorithms that are able to provide continuous feedback as fast and as accurate as possible. To evaluate this aim, the 'steepness' of the time course of the mutual information was used as evaluation criteria and the participants were asked to provide the source code to prove causality.

Despite the requirement to provide the software, 7 participants submitted results. All participants provided some software. In several cases the software could not be tested, because of some missing components. The software was analyzed by visual inspection. In one case an additional delay of 50 samples (0.4s) had to be added.

The winning algorithm is described in [14]. A detailed description of the results is available from [23].

## V. DATA SETS IVA–C: MOTOR IMAGERY

These data sets were provided by Fraunhofer FIRST, Intelligent Data Analysis Group (head: Klaus-Robert Müller), and Charité University Medicine Berlin, Campus Benjamin Franklin, Department of Neurology, Neurophysics Group (head: Gabriel Curio).

### A. Description of data set IVa

All three data sets share the same type of training sessions. Visual cues indicated for 3.5 s which of the following 3 motor imageries the subject should perform: (L) *left* hand, (R) *right* hand, (F) right *foot*. (For IVb and IVc (R) was replaced by (Z) *tongue* (=Zunge in german)). The presentation of target cues were intermitted by periods of random length, 1.75 to 2.25 s in which the subject could relax.

There were two types of visual stimulation: (1) where targets were indicated by letters appearing behind a fixation cross (which might nevertheless induce little target-correlated eye movements), and (2) where a randomly moving object indicated targets (inducing target-uncorrelated eye movements).

Data set IVa poses the challenge of getting along with only a little amount of training data. One approach to the problem is to use information from other subjects' measurements to reduce the amount of training data needed for a new subject. Of course, competitors could also try algorithms that work on small training sets without using the information from other subjects. For this purpose the data sets from five healthy subjects (*aa,al,av,aw,ay*) have been splitted differently into training and test sets, see table IV. Only trials of classes *right* and *foot* were available to the competitors. The performance measure was the overall accuracy. Note that this is not equal to the average across subjects, due to the differently sized test sets. Rather the performance on subjects with large test sets (= small training sets) is weighted stronger.

TABLE IV
THE TOTAL OF 280 TRIALS WAS SPLITTED DIFFERENTLY INTO TRAINING AND TEST FOR EACH SUBJECT. HAVING ONLY A SMALL AMOUNT OF TRAINING SAMPLES POSES A PROBLEM. THIS TABLE SHOWS THE RESPECTIVE NUMBER OF TRAINING (LABELLED) TRIALS (#TRAINING) AND TEST (UNLABELLED) TRIALS (#TEST) FOR EACH SUBJECT.

| subject | #training | #test |
|---|---|---|
| *aa* | 168 | 112 |
| *al* | 224 | 56 |
| *av* | 84 | 196 |
| *aw* | 56 | 224 |
| *ay* | 28 | 252 |

### B. Outcome of the competition – data set IVa

There were 14 submissions for this data set. The winning team is Yijun Wang and collegues from Tsinghua University, Beijing, China. They received accuarcies of 96/100/81/100/98 for the five subjects and an overall accuracy of 94.2 %. This is an excellent performance when considering that the second (Yuanqing Li from the Institute for Infocomm Research, Singapore) and the third best (Liu Yang, National University of Defense Technology, Changsha, Hunan) achieved 85.1 resp. 83.5 %.

The winning team examined three types of features: (1) ERD-feature extracted by Common Spatial Pattern (CSP) analysis, (2) ERD-feature extracted with an AR model, and (3) ERP-feature extracted by LDA on temporal waves. For subjects *aa* and *aw* all three features have been used and combined by a bagging method. For the other 3 subjects only the CSP-based feature was used. To account for the small training sets in subjects *aw* and *ay* a special technique was employed in which formerly classfied test samples are added to the training samples, cf. [15].

### C. Description of data set IVb

Data set IVb poses the problem of classifying in an asynchronous protocol design, i.e., there are no cues indicating that the subject switches to a perdefined mental target class. Rather the subject is by default in an idle state and can spontaneously switch into a mental state that is related to BCI control (here *left* or *foot* imagery). Also the duration of being in that mental state can arbitrarily be decided by the subject. This is in contrast to most classification analyses, which are performed on cued EEG trials, i.e., windowed EEG signals of fixed length, where each trial corresponds to a specific mental state (synchronous protocol). The training data follwed the same experimental setup as in data set IVa. For the competition's

test data set the target classes (*left*, *foot* and *relax*) were ordered by acoustic stimuli in order to have the true labels. The length of those active periods varied between 1.5 and 8 s, intermitted by periods of 1.75 to 2.25 s. The task of the competitors was to give an output signal for each time point of the continuous signals provided as test data. During the intervals of idle state (*relax*) the output is supposed to be small in magnitude (ideally 0), while in periods of *left* resp. *foot* imagery it should be (near to) -1 resp. 1. Note that there are no sample trials for class *relax* in the training data. Rather it has to be defined as absense of the mental states that are used for control. Performance was to be measures by mean square distance of submitted classifier outputs and labels.

### D. Outcome of the competition – data set IVb

Unfortunately, for this data set we received only one submission. So we cannot given an evaluation and elect a winner for this data set. Nevertheless we would like to thank Han Yuan and Yijin Wang from Tsinguhua University very much for their submission. We regret having not receive more submissions for this particular data set, since we think that it poses a highly relevant and difficult challenge.

### E. Description of data set IVc

Data set IVc poses, like IVb, the problem that for a certain amount of test trials the subject was in idle state, i.e., he did not perform motor imagery (class *relax*). The training data for data set IVc is the same as the one for IVb. The experimental setup for the test data was similar to the training sessions, but the motor imagery had to be performed for 1 second only, compared to 3.5 seconds in the training sessions. The length of the intermitting periods ranged from 1.75 to 2.25 seconds as before. The test data was recorded more than 3 hours after the training data, so the distribution of some EEG features could be effected by long-term non-stationarities. The performance criterium is the mean squared error with respect to the target vector that is -1 for class *left*, 1 for *foot*, and 0 for *relax*, averaged across all trials of the test set.

### F. Outcome of the competition – data set IVc

Seven competitors submitted their results to this data set. The winners are Dan Zhang and Yijun Wang from Tsinghua University, Beijing, China. They obtained a mean square error of 0.3 which is much lower than the result of the second best competitor, who achieve 0.59. The different performance becomes explicitly apparent when turning the attention to what the specific challenge of this data set was, the trials of idle state in the test data. These should have been mapped to 0 while left hand and foot motor imagery should have been mapped as -1 and 1 respectively. Fig. 3 shows the histograms of classifier outputs of the two best submissions. Ideally outputs to *left* and *foot* events should all be -1 resp. 1 and outputs to *relax* events (idle state) should be zero. The second best submission performance remarkably well on motor imagery trials but absolutely fails to recognize the idle state trials (as do the other five submissions). The best submission achieves a similar

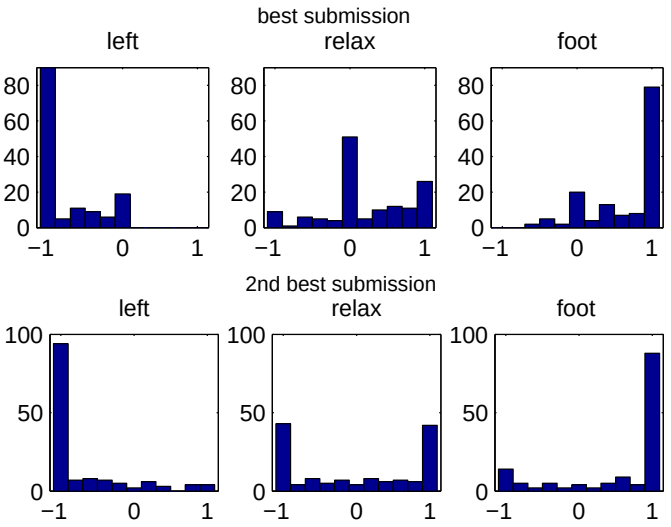


Fig. 3. Histograms of classifier outputs for the two best submissions on data set IVc. Both methods perform well on the motor imagery samples (*left* and *foot*, but only the winning algorithm manages to identify (most of) the idle state samples (*relax*).

good classification of the left and foot imagery events although there are some false negatives. But the particular strength of the method is that it manages to identify more than half of the idle state trials.

The winning team extracted ERD-features by the Common Spatial Subspace Decomposition (CSSD, cf. [27]) method and classified with Fisher Discriminant Analysis. Trials of the *relax* class were detected in a first-pass classification operating on prolonged windows, while the second-pass classified the remaining trials into *left* vs. *foot*, cf. [16] for details.

## VI. Data sets V: multi-class problem, continuous EEG

This data set was provided by the IDIAP Research Institute.

### A. Description of the data set

This data set was recorded from three healthy subjects during four sessions with no feedback. The subject sat in a normal chair, relaxed arms resting on their legs. There are 3 tasks: imagination of repetitive self-paced *left* hand movements, imagination of repetitive self-paced *right* hand movements, and generation of *words* beginning with the same random letter. All 4 sessions of a given subject were acquired on the same day, each lasting 4 minutes with 5-10 minutes breaks in between them. The subject performed a given task for about 15 seconds and then switched randomly to another task at the operator's request. Thus EEG data is not split in trials since the subjects are continuously performing any of the mental tasks. It is worth noting that while operating a brain-actuated application [28], [29], the user does essentially the same as during the recording sessions. The only difference is that in the former case he/she switches to the next mental task as soon as the desired action has been performed, i.e., typically much faster than the 15 s pace in the training sessions.

EEG potentials were measured with a Biosemi portable system using a cap with 32 integrated electrodes located at standard positions of the International 10-20 system. The sampling rate was 512 Hz. Signals were acquired at full DC. No artifact rejection or correction was employed.

Data were provided in two ways, namely, the raw EEG potentials from all 32 electrodes and precomputed features (as described in [30]). The precomputed features were obtained as follows. The raw EEG potentials were first spatially filtered by means of a surface Laplacian. Then, every 62.5 ms —i.e., 16 times per second— the power spectral density (PSD) in the band 8–30 Hz was estimated over the last second of data with a frequency resolution of 2 Hz for the 8 centro-parietal channels C3, Cz, C4, CP1, CP2, P3, Pz, and P4. As a result, an EEG sample is a 96-dimensional vector (8 channels times 12 frequency components).

For each subject there are 3 training files and 1 testing file (the last recording session). The algorithm should provide an output every 0.5 seconds using the last second of data. That is, the goal for the competition was to estimate the class labels for every input vector (either derived from overlapping segments of 1 second of raw EEG data or precomputed sample) of the 3 test files (one per subject).The labels should be estimated in the following way:

1) Precomputed features: Since input vectors are computed 16 times per second, provide the average of 8 consecutive samples (so as to get a response every 0.5 seconds).
2) Raw signals: Compute vectors 16 times per second using the last second of data. Then provide the average of 8 consecutive samples (so as to get a response every 0.5 seconds).

In both cases (precomputed features and raw signals), other (i.e. also past) samples must not be used in order to guarantee a fast response times of the system, although for the competition test data set averaging over more samples could be of benefit. The performance measure is the classification accuracy (correct classification divided by the total number of samples) averaged over the 3 subjects.

### *B. Outcome of the Competition*

There were 26 submissions for this data set, 20 using precomputed features and 6 using raw data. Unfortunately, 4 of the entries did not understand the requirement of using only 1 second of data for estimating the labels and their methods included smoothing consecutive classifier output on longer time windows. Since these results are not comparable to the others, we took them out of the regular scoring. Surprisingly, the best methods used precomputed features. The best submission was by Ferran Galán and colleagues (Univ. of Barcelona) with an error of 31.3 %, but the second-best entry by Xiang Liao (Univ. of Electronic Science and Technology of China) was very close with an error of 31.5 %. In addition, there were 9 contributions with errors between 34.1 % and 40.0 %, of which only one based on raw signals.

## VII. Conclusion and Outlook

Looking at all the winning algorithms of the BCI Competition III reveals several very interesting aspects. (1) Almost all classification methods are linear, which contributes to the linear vs. non-linear debate, cf. [31]. Most popular methods are Fisher Discriminant and linear Support Vector Machines, both introduced in [32] to the field of BCI. (2) In all but one (data set V) cases where multi-channel EEG and oscillatory features were available the winning method used CSSD ([27]) resp. CSP, which was suggested for the use in BCI context in [33]. (3) Several of the winning algorithms incorporated the concept of combining oscillatory (ERD) and non-oscillatory (ERP) features (data sets I, IIIb, IVa), first proposed in [34], [35].

Regarding the distribution of the top performances for each data set we have been astonished by the fact the in all cases except data set V there was a substantial gap between the best and the second best submission, cf. [10]. This is in contrast to the last BCI Competiton, cf. [24], [9] where in most cases the top competitiors had a neck-and-neck race. On the other hand it is interesting to compare the performance achieved on data from different subjects (when available) performing the same mental tasks. In data set IIIa, for example, the best submission achieved an across-subject average kappa value of 0.79 while the least successful submission had a kappa value of 0.64. But on the first of three subjects (K3) the latter submission achieved a very good kappa value of 0.95 where the winner only got 0.82. In data set IIIb the third best team obtained the best result for the first subject (O3) but failed for the second subjects (S4) with a value of 0.09 which is very low compared to 0.44 of the winner. This observation gives raise to the conjecture that brain signals are so specific and divers that specific algorithms are needed. The problem is to select the best suited method given only the training data.

The are some highly relevant topics in BCI research that were not addressed by this competition: (1) transfer of methods and paradigms from offline analyses to feedback applications; (2) optimizing learning in the interaction of two mutually adapting systems human and machine. A complete validation of BCI approachs with regard to those issues within a competition framework would necessitate that all competitors submit real-time versions of their methods which are then tested in a series of online feedback experiments in the hosting BCI laboratories. This could be a new and ambitious objective of a future BCI competition but the effort can be expected to be very high.

The data sets and their descriptions will continue to be available on the competition web page [7]. Other researchers interested in EEG single-trial analysis are welcome to test their algorithms on these data sets and to report their results. To imitate competition conditions, all selections of method, features and model parameters must be confined to the training sets. However, due to the current availability of the labels of the test data and the publication of thorough analyses of these data, future classification results of the competition data cannot fairly be compared to the original submissions.

## Acknowledgment

We thank all people who contributed to this competition, either by submitting classification results, or by giving feedback about the competition.

## REFERENCES


[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, pp. 767–791, 2002.

[2] E. A. Curran and M. J. Stokes, "Learning to control brain activity: A review of the production and control of EEG components for driving brain-computer interface (BCI) systems," *Brain Cogn.*, vol. 51, pp. 326–336, 2003.

[3] A. Kübler, B. Kotchoubey, J. Kaiser, J. Wolpaw, and N. Birbaumer, "Brain-computer communication: Unlocking the locked in," *Psychol. Bull.*, vol. 127, no. 3, pp. 358–375, 2001.

[4] J. del R. Millán, *Handbook of Brain Theory and Neural Networks*, 2nd ed. Cambridge: MIT Press, 2002, ch. Brain-computer interfaces.

[5] P. Sajda, A. Gerson, K.-R. Müller, B. Blankertz, and L. Parra, "BCI competition iii (web page)," 2001. [Online]. Available: http://liinc.bme.columbia.edu/competition.htm

[6] B. Blankertz, "BCI Competition 2003 (web page)," 2003. [Online]. Available: http://ida.first.fhg.de/projects/bci/competition/

[7] ——, "BCI Competition III (web page)," 2004. [Online]. Available: http://ida.first.fhg.de/projects/bci/competition_iii/

[8] P. Sajda, A. Gerson, K.-R. Müller, B. Blankertz, and L. Parra, "A data analysis competition to evaluate machine learning algorithms for use in brain-computer interfaces," *IEEE Trans. Neural Sys. Rehab. Eng.*, vol. 11, no. 2, pp. 184–185, 2003.

[9] B. Blankertz, K.-R. Müller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer, "The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1044–1051, 2004.

[10] B. Blankertz, "BCI Competition III results (web page)," 2005. [Online]. Available: http://ida.first.fhg.de/projects/bci/competition_iii/results/

[11] Q. Wei, F. Meng, Y. Wang, and S. Gao, "BCI Competition III – data set I," *IEEE Trans. Neural Sys. Rehab. Eng.*, 2006, submitted.

[12] A. Rakotomamonjy and V. Guigue, "BCI Competition III – data set II," *IEEE Trans. Neural Sys. Rehab. Eng.*, 2006, submitted.

[13] C. Guan, H. Zhang, and Y. Li, "BCI Competition III – data set IIIa," *IEEE Trans. Neural Sys. Rehab. Eng.*, 2006, submitted.

[14] S. Lemm, "BCI Competition III – data set IIIb," *IEEE Trans. Neural Sys. Rehab. Eng.*, 2006, submitted.

[15] Y. Wang, H. Yuan, D. Zhang, X. Gao, Z. Zhang, and S. Gao, "BCI Competition III – data set IVa," *IEEE Trans. Neural Sys. Rehab. Eng.*, 2006, submitted.

[16] D. Zhang and Y. Wang, "BCI Competition III – data set IVc," *IEEE Trans. Neural Sys. Rehab. Eng.*, 2006, submitted.

[17] F. Galán, F. Oliva, and J. Guàrdia, "BCI Competition III – data set V," *IEEE Trans. Neural Sys. Rehab. Eng.*, 2006, submitted.

[18] T. N. Lal, T. Hinterberger, G. Widman, M. Schröder, J. Hill, W. Rosenstiel, C. E. Elger, B. Schölkopf, and N. Birbaumer, "Methods towards invasive human brain computer interfaces," in *Advances in Neural Information Processing Systems (NIPS) 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 737–744.

[19] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," *IEEE Trans. Biomed. Eng.*, 2004.

[20] E. Donchin, K. M. Spencer, and R. Wijesinghe, "Assessing the speed of a P300-based brain-computer interface," *IEEE Trans. Neural Sys. Rehab. Eng.*, vol. 8, no. 2, pp. 174–179, 2000.

[21] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalogr. Clin. Neurophysiol.*, vol. 70, pp. 510–523, 1988.

[22] A. Schlögl, O. Filz, H. Ramoser, and G. Pfurtscheller, "GDF - a general dataformat for biosignals," 2004. [Online]. Available: http://www.dpmi.tu-graz.ac.at/~schloegl/matlab/eeg/gdf4/TR_GDF.pdf

[23] A. Schlögl, "Results of the BCI-competition 2005 for datasets IIIa and IIIb," 2005. [Online]. Available: http://bci.tugraz.at/schloegl/publications/TR_BCI2005_III.pdf

[24] B. Blankertz, "BCI Competition 2003 results (web page)," 2003. [Online]. Available: http://ida.first.fhg.de/projects/bci/competition/results/

[25] A. Schlögl, C. Neuper, and G. Pfurtscheller, "Estimating the mutual information of an EEG-based Brain-Computer-Interface," *Biomed. Technik*, vol. 47, no. 1-2, pp. 3–8, 2002.

[26] A. Schlögl, "BIOSIG - an open source software library for biomedical signal processing," 2003–2005. [Online]. Available: http://BIOSIG.SF.NET

[27] Y. Wang, P. Berg, and M. Scherg, "Common spatial subspace decomposition applied to analysis of brain responses under multiple task conditions: a simulation study," *Clin. Neurophysiol.*, vol. 110, pp. 604–614, 1999.

[28] J. del R. Millán, F. Renkens, J. Mouriño, and W. Gerstner, "Non-invasive brain-actuated control of a mobile robot by human EEG," *IEEE Trans. Biomedical Engineering*, vol. 51, pp. 1026–1033, 2004.

[29] ——, "Brain-actuated interaction," *Artificial Intelligence*, vol. 159, pp. 241–259, 2004.

[30] J. del R. Millán, "On the need for on-line learning in brain-computer interfaces," in *Proc. Int. Joint Conf. Neural Networks*, 2004.

[31] K.-R. Müller, C. W. Anderson, and G. E. Birch, "Linear and non-linear methods for brain-computer interfaces," *IEEE Trans. Neural Sys. Rehab. Eng.*, vol. 11, no. 2, 2003, 165–169.

[32] B. Blankertz, G. Curio, and K.-R. Müller, "Classifying single trial EEG: Towards brain computer interfacing," in *Advances in Neural Inf. Proc. Systems (NIPS 01)*, T. G. Diettrich, S. Becker, and Z. Ghahramani, Eds., vol. 14, 2002, pp. 157–164.

[33] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filters for single-trial EEG classification in a movement task," *Clin. Neurophysiol.*, vol. 110, pp. 787–798, 1999.

[34] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Combining features for BCI," in *Advances in Neural Inf. Proc. Systems (NIPS 02)*, S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15, 2003, pp. 1115–1122.

[35] ——, "Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 993–1002, June 2004.