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

Benjamin Blankertz, Klaus-Robert Müller, Dean J. Krusienski, *Member, IEEE*, Gerwin Schalk, *Member, IEEE*, Jonathan R. Wolpaw, Alois Schlögl, Gert Pfurtscheller, José del R. Millán, Michael Schröder, and 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. The paper describes the data sets that were provided to the competitors and gives an overview of the results.**

***Index Terms*—Augmentative communication, beta rhythm, brain–computer interface (BCI), electroencephalography (EEG), ERP, imagined hand movements, mu rhythm, nonstationarity, P300, rehabilitation, single-trial classification, slow cortical potentials.**

I. Introduction

Brain–computer interfaces (BCIs) allow their users 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 a microscopic, mesoscopic, or macroscopic scale (cf., [1]–[4] for an overview). The interest in BCI research is rapidly increasing as reflected by the exponentially growing number of published peer-reviewed journal papers on the 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 the internet [5]–[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 labels of the test data were not made public until after the submission deadline, after all submissions were received and evaluated. 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.

TABLE 1
History of BCI Competitions in Numbers: Number of Provided Data Sets, Number of Received Submissions, and Number of Research Labs That Participarted by Submitting Solutions

	date	#datasets	#submissions	#labs
BCI Competition I	2001/2002	3	10	8
BCI Competition II	2002/2003	6	57	20
BCI Competition III	2004/2005	8	92	49

A. History of BCI Competitions

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; see Table I. The tasks and results of the first two competitions are summarized in [8] and [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 important analysis challenges that are highly relevant to present BCI research.

More specifically, the competition comprised the problems of session-to-session transfer, nonstationarity, small training sets, subject-to-subject transfer, continuous test data without trial structure, asynchronous paradigms, and idle states (see Table II for an overview).

B. 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 for a binary problem, each of which is able to classify 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).

Manuscript received July 19, 2005

The work of B. Blankertz and K.–R. Müller was supported in part by the Bundesministerium für Bildung und Forschung (BMBF), under Grant FKZ 01IBE01A/B and by the Deutsche Forschungsgemeinschaft (DFG), under Grant FOR 375/B1. The work of D. J. Krusienski, G. Schalk, and J. R. Wolpaw was supported in part by the National Institutes of Health under Grants HD30146 (National Center for Medical Rehabilitation 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 by the James S. McDonnell Foundation. The work of J. d. R. Millán was supported by the Swiss National Science Foundation under Grant NCCR "IM2." The work of B. Blankertz, K.-R. Müller, and J. d. R. Millán was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, under Grant IST-2002-506778.

B. Blankertz is with Fraunhofer FIRST (IDA), D-12489 Berlin, Germany (e-mail: benjamin.blankertz@first.fraunhofer.de).

K.-R. Müller is with Fraunhofer FIRST (IDA), D-12489 Berlin, Germany, and also with the University of Potsdam, 14415 Potsdam, Germany.

D. J. Krusienski and G. Schalk are with the Laboratory of Nervous System Disorders, Wadsworth Center, New York State Department of Health, Albany, NY 12208 USA.

J. R. Wolpaw is with the Laboratory of Nervous System Disorders, Wadsworth Center, New York State Department of Health, Albany, NY 12208 USA and also with the State University of New York, Albany, NY 12222 USA.

A. Schlögl and G. Pfurtscheller are with the Institute for Human-Computer Interfaces, University of Technology Graz, A-8010 Graz, Austria.

J. d. R. Millán is with the IDIAP Research Institute, CH-1920 Martigny, Switzerland.

M. Schröder is with the Department of Technical Computer Science, Eberhard-Karls-Universität Tübingen, 72076 Tübingen, Germany.

N. Birbaumer is with the Institute of Medical Psychology and Behavioral Neurobiology, University of Tübingen, 72076 Tübingen, Germany, and also with the University of Trento, 38100 Trento, Italy.

Digital Object Identifier 10.1109/TNSRE.2006.875642

TABLE II
OVERVIEW OF DATA SETS OF BCI COMPETITON III. MOST DATA SETS WERE RECORDED IN MOTOR IMAGERY PARADIGMS, EXCEPT FOR DATA SET II WHICH WAS RECORDED IN P300 SPELLER APPLICATION

#	lab	#channels	challenge
I	Tübingen	64 ECoG	distribution shift: training → test set
II	Albany	64 EEG	detection of P300 in fair subject
IIIa	Graz	60 EEG	multi-class, quite good and fair subjects
IIIb	Graz	2 EEG	non-stationarity EEG signals
IVa	Berlin	118 EEG	small training sets
IVb	Berlin	118 EEG	uncued test data with various epoch length
IVc	Berlin	118 EEG	test data include a 'no action' class
V	Martigny	32 EEG	multi-class, uncued test data

TABLE III
WINNING TEAMS FOR ALL COMPETITION DATA SETS ARE LISTED. SECTION V INCLUDES 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

C. Overview of Paper

In Sections 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 III.

II. DATA SET I

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

A. Description of 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 nonstationarity 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. As electrocorticography (ECoG) was used and not electroencephalography (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.

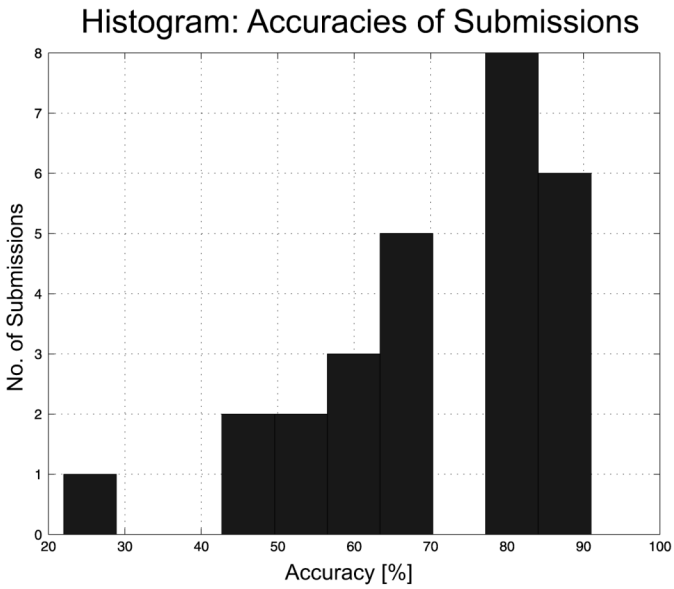

Fig. 1 Histogram of classification accuracy of 27 submitted solutions. One submission stays clearly below chance level of 50%. Group of 14 submissions reaches more than 78% accuracy.

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 8×8 ECoG platinum electrode grid which was placed on the contralateral (right) motor cortex. The grid was covered by meninges and skull and was not sensitive to muscle artifacts. As the skull 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 (approximately 8×8 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 3 s. To avoid visually evoked potentials being reflected by the data, the recording intervals started 0.5 s after the visual cue had ended. For further information about the experiment, please refer to [11].

B. Outcome of 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% classification accuracy on the test set. Although including an outlier of only 22% accuracy (probably submitted with accidentally confused class labels), the average accuracy of all submissions was 70%. Fig. 1 shows the histogram of the submission accuracy.

The submissions of rank one to three and their applied methods at a glance are as follows.

1) An accuracy of 91% was achieved by Q. Wei and his cocontributors from the Tsinghua University, Beijing, China. They used a

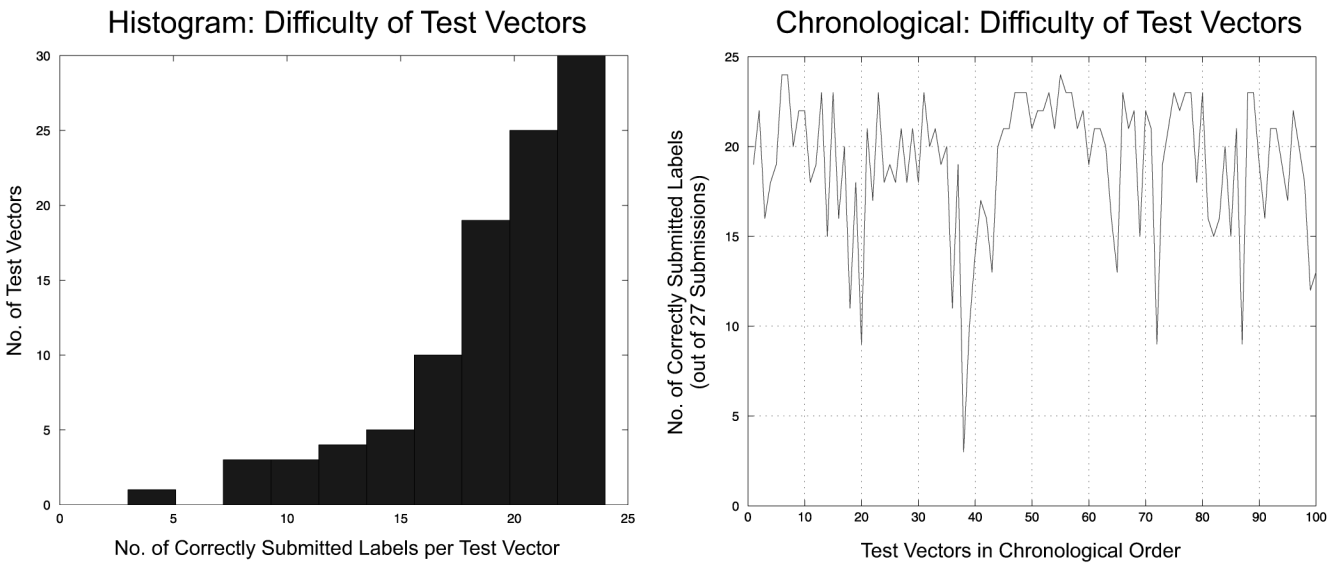

Fig. 2 Difficulty of test vectors from contributor's point of view. Left histogram shows no vector was misclassified by every submission and many vectors received correct labels from 20 or more submissions. Another view of distribution provides right graph. It shows number of correctly submitted labels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

combination of bandpower features together with common spatial subspace decomposition (CSSD) and mean waveforms that were chosen by Fisher discriminant analysis before classification was performed with a linear support vector machines (SVM).

2) An accuracy of 87% was achieved by P. Hammon from the University of California, San Diego. After unmixing with independent component analysis (ICA), a combination of coefficients of autoregressive (AR) models, spectral power (0-45 Hz), and wavelet coefficients were used as features. Classification was performed with regularized logistic regression.
3) Marginally less, 86% accuracy, was reached by three submissions: By M. Sapinski from Warsaw University, by M. Dawei and cocontributors from Zhejiang University, and by A. D'yakonov from Moscow State University. Their features comprise the offset and spectral power of hand selected channels (M. Sapinski), the standard deviation of the Hilbert–Huang Transform for time frequency windows (window size: 5 Hz and 0.2 s) of seven channels (M. Dawei), and hand chosen features from seven channels (A. D'yakonov). For classification, logistic regression (M. Sapinski) and Mahalanobis distance (M. Dawei) were used.

Taking a closer look on solutions above 60% accuracy, discriminant analysis (linear, robust, etc.) dominates the classification methods with four entries, and (linear) support vector machines have three 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 at 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 nonstationarity in the signals caused by epileptiform patterns in the EEG which did arise frequently for this patient.

III. Data Set II: P300 Speller Paradigm

This data set was provided by the BCI Laboratory of the Wadsworth Center, New York State Department of Health (Head: J. R. Wolpaw).

A. Description of Data Set

This data set represents a complete record of P300 evoked potentials (five sessions from two subjects) recorded with the BCI2000 software [12], using a paradigm described in [13] and originally by Farwell and Donchin [14]. 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 six rows and six 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 two 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 [13], [14]. 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. The time between intensifications was 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, during which the matrix was displayed. This period informed the user that the character was completed and that s/he should focus attention on the next character in the word that was displayed on the top of the screen (the 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 five flash sequences.

TABLE IV
MAXIMUM KAPPA FOR $t \leq 7s$ IN THREE SUBJECTS (K3, K6, L1) AND ITS MEAN OBTAINED BY 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

B. Outcome of Competition

A total of ten submissions were received for this data set, incorporating a wide variety of preprocessing and classification methods. Using all 15 sequences, the majority of submissions (eight) 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%.

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: G. Pfurtscheller).

A. Description of Data Set IIIa

The data set consists of recordings from three subjects; the subjects performed four 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 at 250 Hz and was filtered between 1 and 50 Hz with the notchfilter on. The data of all runs was concatenated and converted into the general dataformat (GDF), cf. [15]. 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 consisted of several runs (at least six) with 40 trials each; after the trial began, the first 2 s were silent, at $t = 2$ s an acoustic stimulus indicated the beginning of the trial, and a cross "+" was displayed. Then, from $t = 3$ s, an arrow to the left, right, up, or down was displayed for 1 s; 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=7 s. Each of the four cues was displayed ten times within each run in a randomized order. The participants needed to provide a continuous classification output (continuous in time as well as magnitude) for all four classes. In other words, the classifier needed to provide four 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 determined the corresponding class. Then, a confusion matrix was built from all trials for each time-point 0.0 s $\leq t \leq$ 7.0 s. From these confusion matrices, the time course of the accuracy and the time-course of the kappa coefficient was obtained. The performance measure of the competition entry was the maximum kappa value in time, averaged for the three subjects.

B. Outcome of Competition — Data Set IIIa

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

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

TABLE V
MAXIMUM STEEPNESS (WITH $t_0 = 3$ S) OF MUTUAL INFORMATION [BITS/S] IN THREE SUBJECTS (O3, S4, X11) AND ITS MEAN OBTAINED BY 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

2) Authors Guan, Zhang, and 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 common spatial pattern (CSP), SVM.
3) Authors Gao (Head), Wu, Wei (Tsinghua University, Beijing, China); Method: surface laplacian, 8–30-Hz filter, CSP (one-versus-rest), SVM+kNN+LDA, "bagging."

A detailed description of the results is available from [16].

C. Description of Data Set IIIB

Data set IIIb contained two-class EEG data from three subjects. Each data set contained recordings from consecutive sessions during a BCI experiment. The large amount of data enabled the use of nonstationary classifiers. Time-varying classifiers are expected to perform better than stationary (static) classifiers. Moreover, based on the experience of the second BCI competitition [6]–[17], the response time of each method has to be evaluated. The experiment consists of three sessions for each subject. Each session consists of four to nine runs. The data of all runs was concatenated and converted into the GDF format [15]. The recordings were made with a bipolar EEG amplifier from g.tec. The EEG was sampled at 125 Hz and was filtered between 0.5 and 30 Hz with the 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 algorithms used were causal. The output was validated using the time course of the mutual information [18]. 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 >$3.5 s was evaluated. The "steepness" of the mutual information quantified the response time. The evaluation algorithm is provided in BIOSIG (see /biosig/t490/criteria2005IIIb.m in [19]).

D. Outcome of 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 V.

1) Authors O. Burmeister, M. Reischl, and R. Mikut (Forschungszentrum Karlsruhe, Germany); Method: Bandbower (BP), ratios and differences of BP; MANOVA for feature selection; SVM and linear combiner.
2) Author S. Lemm (Fraunhofer-FIRST IDA, Berlin, Germany); Method: event-related potential (ERP) and event-related desynchronization (ERD) features (mu and beta), propabilistic classification model, accumulative classifier.
3) Authors X. Pei and G. Bin (Institute of Biomedical Engineering of Xi'an Jiaotong University, Xi'an, China); Method: Fast Fourier transform (FFT) with Hanning window of 1-s segments; Fisher discriminant analysis.

The main aim was to evaluate causal algorithms that are able to provide continuous feedback that is fast and accurate. To evaluate this aim,

TABLE VI
TOTAL OF 280 TRIALS WAS SPLIT DIFFERENTLY INTO TRAINING AND TEST FOR EACH SUBJECT. HAVING ONLY A SMALL AMOUNT OF TRAINING SAMPLES POSES A PROBLEM. 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

the "steepness" of the time course of the mutual information was used as the evaluation criterion and the participants were asked to provide the source code to prove causality.

Despite the requirement to provide the software, seven 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 (0.4 s) had to be added. A detailed description of the results is available from [16].

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: G. 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 three 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 was 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) targets were indicated by letters appearing behind a fixation cross (which might nevertheless induce little target-correlated eye movements) and 2) 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*) were split into training and test sets (see Table VI). 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 (equating to small training sets) is weighted stronger.

B. Outcome of Competition — Data Set IVa

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

The winning team examined three types of features: 1) ERD feature extracted by 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 three 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 were added to the training samples.

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 predefined 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 followed 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 zero), while in periods of *left* (respectively *foot*) imagery it should be (near to) -1 (respectively, 1). Note that there are no sample trials for class *relax* in the training data. Rather, it has to be defined as absence of the mental states that are used for control. Performance was to be measured by mean square distance of submitted classifier outputs and labels.

D. Outcome of Competition — Data Set IVb

Unfortunately, for this data set we received only one submission. So, we cannot give an evaluation and elect a winner for this data set. Nevertheless, we would like to thank H. Yuan and Y. Wang from Tsinguhua University very much for their submission. We regret having not received 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 s only, compared to 3.5 s in the training sessions. The length of the intermitting periods ranged from 1.75 to 2.25 s as before. The test data was recorded more than 3 h after the training data, so the distribution of some EEG features could be effected by long-term nonstationarities. The performance criterion 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 Competition — Data Set IVc

Seven competitors submitted their results to this data set. The winners were D. Zhang and Y. Wang from Tsinghua University, Beijing, China. They obtained a mean square error of 0.3, which was much

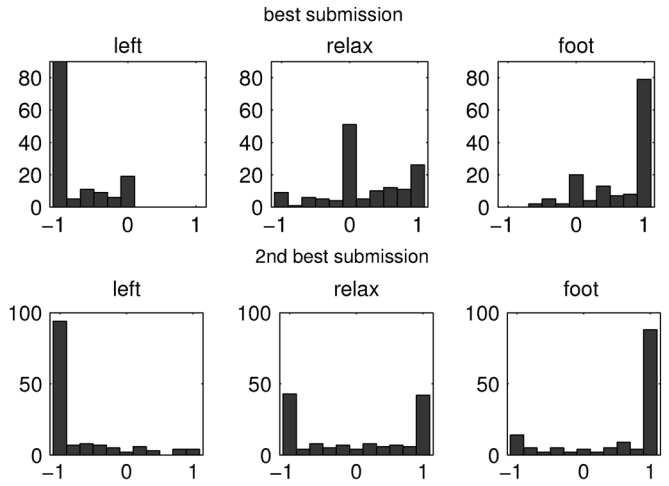

Fig. 3 Histograms of classifier outputs for two best submissions on data set IVc. Both methods perform well on motor imagery samples (*left* and *foot*), but only the winning algorithm manages to identify (most of) idle state samples (*relax*).

lower than the result of the second best competitor, who achieved 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 zero 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 (respectively, 1) and outputs to *relax* events (idle state) should be zero. The second best submission performed remarkably well on motor imagery trials but absolutely failed to recognize the idle state trials (as did the other five submissions). The best submission achieved a similarly good classification of the *left* and *foot* imagery events although there are some false negatives. But, the particular strength of the method was that it managed to identify more than half of the idle state trials.

The winning team extracted ERD features by the CSSD (cf., [20]) 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* versus *foot*.

VI. Data Sets V: Multiclass Problem, Continuous EEG

This data set was provided by the IDIAP Research Institute (Head: J. del R. Millán).

A. Description of Data Set

This data set was recorded from three healthy subjects during four sessions with no feedback. The subject sat in a normal chair, with relaxed arms resting on their legs. There were three 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 four sessions of a given subject were acquired on the same day, each lasted 4 min with 5–10-min breaks between them. The subject performed a given task for about 15 s and then switched randomly to another task at the operator's request. Thus, the EEG data were not split into trials since the subjects were always performing one of the mental tasks. It is worth noting that while operating a brain-actuated application [21], [22], the user does essentially the same as during the recording sessions. The only difference is that in the former case, s/he 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 [23]). 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 eight centro-parietal channels C3, Cz, C4, CP1, CP2, P3, Pz, and P4. As a result, an EEG sample is a 96-dimensional vector (eight channels times 12 frequency components).

For each subject, there are three training files and one testing file (the last recording session). The algorithm should provide an output every 0.5 s 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 one second of raw EEG data or precomputed sample) of the three test files (one per subject).The labels needed to be estimated in the following way.

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

In both cases (precomputed features and raw signals), other (i.e., also past) samples could 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 three subjects.

B. Outcome of Competition

There were 26 submissions for this data set, 20 using precomputed features and six using raw data. Unfortunately, four of the entries did not understand the requirement of using only 1-s 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 F. Galán and colleagues (University of Barcelona) with an error of 31.3%, but the second best entry by X. Liao (University of Electronic Science and Technology of China) was very close with an error of 31.5%. In addition, there were nine contributions with errors between 34.1% and 40.0%, of which only one was 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 versus nonlinear debate (cf., [24]). The most popular methods are Fisher discriminant [25], [26] and linear SVM [26]. 2) In all but one case (data set V) where multichannel EEG and oscillatory features were available, the winning method used CSSD ([20]) /CSP, which was suggested for the use in BCI context in [27]. 3) Several of the winning algorithms incorporated the concept of combining oscillatory (ERD) and nonoscillatory (ERP) features (data sets I, IIIb, IVa), proposed in [28]–[30].

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., [17] and [9]) where in most cases the top competitiors were very close in performace. 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 and 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 subject (S4) with a value of 0.09, which is very low compared to 0.44 from the winner. This observation gives rise to the conjecture that brain signals are so specific and diverse across individuals that specific algorithms are needed. The problem is to select the best 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 and 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 require that all competitors submit real-time versions of their methods for testing 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

The authors would like to thank all people who contributed to this competition, either by submitting classification results or by giving feedback about the competition. This publication only reflects the authors' views.

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, MA: MIT Press, 2002.

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

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

[7] B. Blankertz, "BCI Competition III," , 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 Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 184–185, Mar. 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, Jun. 2004.

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

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

[12] 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.*, no. 6, pp. 1034–1043, Jun. 2004.

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

[14] 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.

[15] 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

[16] 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

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

[18] 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.

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

[20] 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. Neurophys.*, vol. 110, pp. 604–614, 1999.

[21] 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. Biomed. Eng.*, vol. 51, no. 6, pp. 1026–1033, Jun. 2004.

[22] J. del R. Millán, F. Renkens, J. Mouriño, and W. Gerstner, "Brain-actuated interaction," *Artif. Intell.*, vol. 159, pp. 241–259, 2004.

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

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

[25] L. A. 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, no. 6, pp. 510–523, Dec. 1988.

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

[27] 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.

[28] G. Schalk, J. R. Wolpaw, D. J. McFarland, and G. Pfurtscheller, "EEG-based communication: Presence of an error potential," *Electroencephalogr. Clin. Neurophysiol.*, vol. 111, no. 12, pp. 2138–2144, Dec. 2000.

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

[30] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "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, Jun. 2004.