Multimodal Classification of EEG During Physical Activity

Yi Ding
yding@cs.ucsb.edu
University of California Santa Barbara

Brandon Huynh
bhuynh@cs.ucsb.edu
University of California Santa Barbara

Aiwen Xu
aiwenxu@cs.ucsb.edu
University of California Santa Barbara

Tom Bullock
thomas.bullock@psych.ucsb.edu
University of California Santa Barbara

Hubert Cecotti
hcecotti@mail.fresnostate.edu
California State University, Fresno

Matthew Turk
mturk@ucsb.edu
University of California Santa Barbara

Barry Giesbrecht
barry.giesbrecht@psych.ucsb.edu
University of California Santa Barbara

Tobias Höllerer
holl@cs.ucsb.edu
University of California Santa Barbara

ABSTRACT

Brain Computer Interfaces (BCIs) typically utilize electroencephalography (EEG) to enable control of a computer through brain signals. However, EEG is susceptible to a large amount of noise, especially from muscle activity, making it difficult to use in ubiquitous computing environments where mobility and physicality are important features. In this work, we present a novel multimodal approach for classifying the P300 event related potential (ERP) component by coupling EEG signals with nonscalp electrodes (NSE) that measure ocular and muscle artifacts. We demonstrate the effectiveness of our approach on a new dataset where the P300 signal was evoked with participants on a stationary bike under three conditions of physical activity: rest, low-intensity, and high-intensity exercise. We show that intensity of physical activity impacts the performance of both our proposed model and existing state-of-the-art models. After incorporating signals from nonscalp electrodes our proposed model performs significantly better for the physical activity conditions. Our results suggest that the incorporation of additional modalities related to eye-movements and muscle activity may improve the efficacy of mobile EEG-based BCI systems, creating the potential for ubiquitous BCI.

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3353759

CCS CONCEPTS

• **Human-centered computing** → **Interaction techniques**; **Empirical studies in HCI**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Machine Learning, Neuroscience, Electroencephlogram, EEG, Brain Computer Interfaces, Dataset, Motion, Deep Learning, Neuroadaptive Technology

ACM Reference Format:
Yi Ding, Brandon Huynh, Aiwen Xu, Tom Bullock, Hubert Cecotti, Matthew Turk, Barry Giesbrecht, and Tobias Höllerer. 2019. Multimodal Classification of EEG During Physical Activity. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340555.3353759

1 INTRODUCTION

Brain Computer Interface (BCI) systems enable the control of a computer through brain signals [49]. Traditionally, BCIs have been utilized as an assistive technology for people with mobility impairments [44]. There is however a growing interest in general purpose, non-invasive BCI technologies to improve the computing experience of perfectly healthy people. A number of consumer facing products have been developed such as EEG headsets by Emotiv, the Muse meditation headband, and an EEG-integrated virtual reality headset by Looxid Labs. These products promise to enhance the computing experience by enabling intelligent interfaces that sense and react to changes in the user's cognition.

Recent work from neuroadaptive systems have explored the use of these brain signals (EEG) as an additional input modality for a wide variety of interaction tasks. A central idea is to use EEG information for user modeling to build adaptive interfaces [18, 19] that can implicitly and quickly

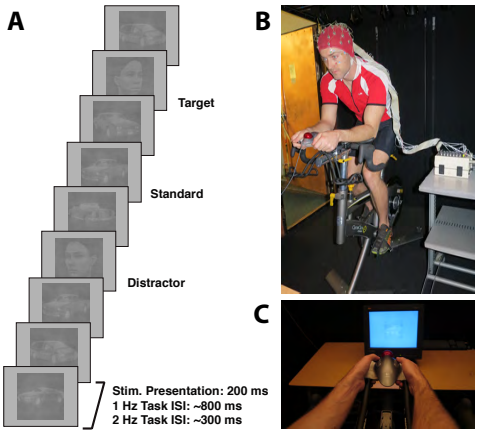

Figure 1: Methods and Tasks. (A) Example of the oddball task. Participants were required to detect targets (right oriented faces) in a stream of distractors (left oriented faces) and standards (cars oriented left or right). (B) The participant was fitted with an EEG cap and positioned on a stationary bike. (C) The participant rested their elbows on a pair of "aero bars" attached to the bike handlebars and used their right thumb to respond to targets.

react to user state [26, 41]. This information can be used to quantify user states such as cognitive load [20, 27], emotion [4, 25], and attention [15, 31] to inform better interaction experiences for education [30], entertainment [29], equipment operation [52], and others [2, 3, 32, 45]. Nonetheless, an open research question remains as to how these signals should be integrated and how reliable they are for non-trivial computing applications.

The key advantage of these technologies is that they opens the door to real Ubiquitous Computing, where computing may occur in any time or place [1]. In ubiquitous computing, users may be interacting with the system while moving around in their environment and engaging in physical activity (e.g. an augmented reality task). However, EEG signals are commonly known to be severely impacted by a wide range of biophysiological artifacts associated with movement. To ensure that the system remains usable, it is crucial to understand how classification performance of EEG signals changes under these adverse conditions, so that we can develop techniques for robust classification and analysis.

Typical BCI solutions are based on laboratory studies and rarely replicate the conditions outside the lab in which the system should be deployed. There are three main ways to tackle such a problem: 1) to extract features that are invariant to the expected noise, 2) to denoise the signal, and 3) to be robust to the noise. An extensive amount of data collection, manual feature extraction, and domain knowledge is typically necessary to identify, classify, and correlate these signals to a particular application [6]. A number of techniques have been used to alleviate the need for manual feature extraction, including spatial and temporal filtering, and neural networks [33].

Due to the success of Deep Learning models in other fields such as Computer Vision and Speech Recognition in which the performance reaches human-like levels of performance, there has been a resurgence in the use of deep neural networks for feature extraction and classification of EEG signals [43]. However, deep learning methods require a large amount of training examples to be successful in order to model the variability that exists across examples [23], which is not typically the case for BCI data. First, the EEG datasets have a low number of examples per class compared to typical computer vision problem, and they have unbalanced datasets, in particular for event-related potential (ERP) based BCI in which the target class has a low probability. Second, EEG is highly non-stationary and characteristics of the signal can change depending on the behavioral state of the wearer (e.g. fatigue/arousal). These effects can be partially remedied through the use of data augmentation [14, 39, 46], but the applications of these techniques have not been well studied for EEG signals.

To help address these issues, we introduce a dataset in which participants were positioned on a stationary bike and engaged in a visual three-stimulus oddball task [38] while at rest and during bouts of low- and high- intensity cycling exercise [9]. We open source[1] this dataset with the goal of encouraging further research. We investigate whether classification performance of a state-of-the-art deep learning model suffers under different intensity levels of physical activity, and discover that it does, suggesting room for improvement of feature representation. We propose a model to improve performance by incorporating the use of a denoising autoencoder. Furthermore, we consider the addition of signals from nonscalp electrodes and user state data, to provide supplementary information.

The paper is organized as follows. First, the related work and classification methods are described in Section 2. Second, the datasets are detailed in Section 3. Third, the methods are presented in Section 4. Finally, the classification results and the impact of the proposed method are discussed in Section 6.

2 RELATED WORK

Classification Methods

Lotte et al. [33, 34] has provided a review of classification algorithms for EEG-based BCI. They concluded that the current state-of-the-art is Riemannian Geometry (RG) classifiers, and suggested it is time to move away from classical approaches which usually use Linear Discriminant Analysis (LDA) with Common Spatial Pattern (CSP) filters. In fact, the winning

[1] https://github.com/yding37/mcann

approach for the Kaggle BCI competition at NER 2015 used xDAWN spatial filters with RG [5].

Deep learning models have been applied to EEG classification since at least 2008 [12] and there has been a sharp increase in activity thanks to the recent success of these models in the natural language processing and computer vision domains. The best performing deep learning approach is currently Convolutional Neural Networks (CNN), which are able to borrow technical advancements from the computer vision community. CNNs were first used for EEG classification by Cecotti et al. in 2011 [11, 13] for P300 ERP classification. Schirrmeister et al. [43] conducted an in-depth survey of CNN architectures for EEG classification and provided an open source software library for evaluating them.

Recently, the US Army Research Lab released EEGNet, a CNN architecture that reached performance comparable to the state-of-the-art on 4 different BCI tasks [28]. The authors used depthwise and separable convolutions which helped to reduce the amount of trainable parameters in the model [24]. Our model uses a variant of EEGNet as the basis for our encoder. By applying EEGNet within an autoencoder paradigm, we are able to learn a more robust representation of the EEG data.

Although not as common, autoencoder methods have been explored in a few EEG classification studies. In [51], the authors utilized a Stacked Denoising Autoencoder (SDEA) in order to classify mental workload. We also utilize a denoising autoencoder, but we target the P300 signal which is better characterized and understood [37]. The authors also compared computation time and concluded that their SDEA model could be used for online classification, which supports our intended use case of ubiquitous BCI.

[42] used a multimodal fusion approach with stacked autoencoders coupling EEG and EMG data. However, their approach utilized two separate pathways for EEG and EMG and learned a latent vector representation of the data. However, their network did not show improvements over CNN based approaches. In our approach, we share the weights of the encoder for both EEG and NSE and fuse their latent representations.

EEG During Physical Activity

EEG recordings are known to suffer from motion artifacts, as they simply measure the electrical signals in the brain. Moreover, the activation of muscles produces large electrical signals and is the main source of noise in many studies. Participants are usually trained to stay completely still in order to minimize contamination of the dataset. For this reason, there are relatively few datasets where EEG is actually recorded while under motion.

In [36] the authors collected EEG recordings of subjects walking at 3 different speeds on a treadmill. Contrary to expectations, they did not observe significant contamination of the EEG signal by motion artifacts. However, it should be noted that the fastest speed investigated (4.5 km/h) is still less than the preferred walking speed of an average person [8], making it difficult to extrapolate to Ubiquitous BCI settings.

Table 1: Our dataset is the largest among current publicly available datasets for BCI tasks of similar purpose.

Dataset	# Samples	Ratio
BCI-2a [47]	2.5K	1:1
Kaggle [35]	8.8K	1:1
P300 [28]	30K	1:5.6
Bike (ours) [9]	72K	1:1:8

A handful of studies have collected EEG data during acute bouts of exercise. Yagi et al. [50] and Grego et al. [21] both measured EEG with a P300 task while cycling. More recently, other studies have investigated the impact of acute exercise on other types of brain responses, such as orientation-selective responses in visual cortex [10] and neural oscillatory activity associated with inhibitory control [17]. For a comprehensive summary of sport and exercise related EEG studies, see Cheron et al. [16]. However, to the best of our knowledge, the present study is the first to apply classification methods to EEG collected during acute bouts of physical activity, with the goal of improving the efficacy of ubiquitous BCI.

Artifact removal is another common practice for removing sources of interference such as muscle activity. Gwin et al. [22] compare artifact removal methods for EEG collected while walking or running. General guidelines and good practice for artifact removal can be found in [48]. Ultimately, given that we are interested in online classification, we chose not to perform any motion artifact removal. Instead, we perform minimal pre-processing on the datasets, which we describe in the next section.

3 DESCRIPTION OF DATASET

Task and Exercise Protocol

The EEG dataset used in this work was previously described in [9]. Twelve adult student volunteers took part in the study in exchange for course credit or financial compensation. Figure 1 provides an overview of the methodology used for collection. Each participant performed two different versions of a three-stimulus oddball task [38] while seated on a stationary bike. Participants were required to respond to target stimuli (left-oriented faces) and ignore the distractor stimuli (right-oriented faces) and the standard stimuli (cars oriented either to the left or right). The ratio of targets to distractors and standards was 1:1:8, respectively. In the two different versions of the task the stimuli were presented at different rates. Stimuli were either presented at 1 Hz (200 ms stimulus

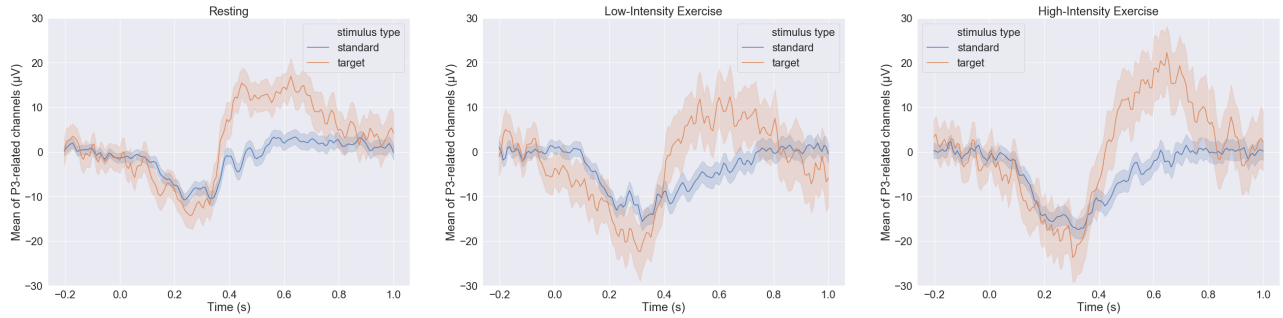

Figure 2: Single subject (sj04) ERPs are shown for each of the physical activity conditions. To avoid visual clutter, only ERPs generated from the standard and target conditions are shown. Error bars represent Âś standard error of the mean (SEM)

presentation with 800 ms inter-stimulus intereval (ISI) or 2 Hz (200 ms stimulus presentation with 300 ms ISI). The 2 Hz data were collected for the purpose of a BCI study and were not reported in the original paper.

Participants completed the 1 Hz and 2 Hz tasks at rest (sat on the bike but not pedaling), during low-intensity exercise (pedaling at at a very light resistance level of 40W) and during high-intensity exercise (pedaling at a resistance level which the participant reported to be "somewhat hard" according to their Rating of Perceived Exertion (RPE; Borg 1970 [7].)). The order of completion was counterbalanced between participants.

EEG data were recorded continuously during each task using a BioSemi Active Two System consisting of 32 scalp electrodes arranged in an elastic cap (Electro-Cap, OH, USA) and placed in accordance with the 10-20 system. Additional non-scalp electrodes (NSE) were fixed to the right and left mastoids, 1 cm lateral to the left and right canthi (horizontal EOG), above and below each eye (vertical EOG) and on the right and left trapezius muscles (EMG).

Classification Goals and Challenges

Here, the goal of the classifier was to determine which stimulus (target, distractor or standard) the participant viewed for each trial. The inclusion of two physical activity conditions sets this dataset apart from typical P300 datasets. This dataset fits with our goal of building BCI paradigms for ubiquitous computing, because compared to other P300 datasets, the conditions in this task are more similar to those that might be encountered in real life. In traditional P300 EEG data collection, participants typically sit in a comfortable position and are told to minimize non-task related physical motion. However, for BCIs to be useful in real life, the actions of the user should not be controlled. In contrast, this P300 dataset incorporates physical exercise, an indispensable part of day-to-day life. Therefore, achieving a good classification accuracy on this dataset is a first step to building a useable BCI for everyday activities.

The inclusion of physical exercise introduced extra noise, which makes the classification task more difficult. The presence of extra noise under physical exercise is visualized in the error bands in Figure 2. Here, we identify at least three sources of noise which may not be present in other existing P300 datasets. The first is EMG noise. EMG activity can be present in the range 10 - 250 Hz, which overlaps with the useful frequency band of EEG signals at 1 - 40 Hz. The second is that sweating can cause low-frequency noise [40]. The third is physical motion itself. During the task, the participants were biking at 50 RPM, which corresponds to 100 pedal downstrokes per minute (left and right). This introduced a noise at 1.33 Hz, which also overlaps with the 1 - 40 Hz range. Due to the overlap, naive filtering methods cannot eliminate those sources of noise completely.

4 METHOD

The ERP classification task is defined as follows: A set of EEG channels C and its signal over time $\mathbf{x} \in \mathbb{R}^{C \times T}$ is given where T depends on the sampling rate and duration of an epoched trial. The task is to take each epoched trial x and output a 3-class probability distribution $\mathbf{y}$. We are additionally given $\mathbf{x}_{nse} \in \mathbb{R}^{C_n \times T}$ for NSE information and $\mathbf{s} \in [0.1, 0.5, .9]$ for resting, low, and high exercise states.

In this paper, we propose an end-to-end deep learning architecture, Multimodal Context-Aware Neural Network (MCANN), for modeling the ERP prediction problem. Our model (Figure 3) is in part motivated by the ability for unsupervised techniques to build a good representation of data. We break our model into 4 components: 1) a temporal feature extraction module in section , 2) a fusion component which combines these features, 3) a decoder which to reconstruct the signal for unsupervised learning and 4) a classification network for predicting the final class distribution.

Unsupervised Representation Learning

Autoencoders first map an input $\mathbf{x}$ into a latent representation by a deterministic mapping: $\mathbf{z} = f_\theta(x)$. The latent

Table 2: Temporal Encoding Network

Layer	Parameters
Conv2d	1x10x10
Batch Norm	f: 10, eps: 1e-3, m:0.1
Conv2d, Elu	Cx10x10
AvgPool	k: 1x4 s: 1x4
Renorm	p: 2, mn: 1
Conv2d	10x1x20

Table 3: Fusion Network

Layer	Parameters
Dropout	p: .25
Conv2d	1x16x20, g: 20
Conv2d	1x1x10
AvgPool	k: 1x8, s: 1x8
Dropout	p: .25
Fully Connected, Elu	(T/1.6)x64

Table 4: Decoder Network

Layer	Parameters
Fully Connected, Elu	64x(20*T//32)
BatchNorm	
Deconv2d	1x1x20
Fully Connected, Elu	(T/4 + 1)x(T/4 + 1)
Deconv2d	Cx1x10, g: 10
Deconv2d, Elu	1x5x1
Deconv2d	1x10x1

***Network parameters* are abbreviated as follows: eps for epsilon, m for momentum, f for number of filters, k for kernel size, s for stride, p for power, mn for max norm, and g for groups. Convolutions filter sizes are expressed in channel by time by number of filters.**

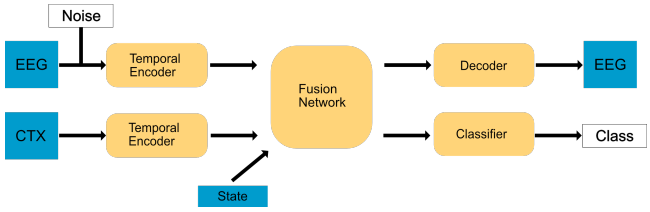

Figure 3: Our proposed model for evaluating EEG data with additional input modalities.

representation $\mathbf{z}$ is then remapped back to $\mathbf{x}' = g_{\theta'}(\mathbf{z})$. A denoising autoencoder additionally takes a corrupted version of the original input $\tilde{\mathbf{x}}$ to reconstruct $\mathbf{x}$. Autoencoders of this sort have been shown to be robust to partial destruction of input for a wide range of tasks. Here $f_\theta(x)$ and $g_\theta(x)$ are modeled by multiple neural networks.

Noisy input $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{Wn}$ is formed by the addition of a noise vector whose values are sampled independently from a normal distribution $\mathbf{Wn}_{ij} \sim \mathcal{N}(0, \sigma^2)$ for all $i \in \{1, \ldots, C\}$ channels and all $j \in \{1, \ldots, T\}$ time samples. Here we use the standard normal distribution ($\sigma^2 = 1$), however other values could be considered depending on the dataset. Additionally, alternative methods of corruption could be explored which can be informed via user context or state information.

The noisy input $\widetilde{\mathbf{x}}$ is concatenated with prior noise information $\mathbf{x}_{nse}$ and $\mathbf{s}$ and fed into our network. The reconstructed signal $\mathbf{x}' = g_{\theta'}(f_\theta(\tilde{\mathbf{x}}, \mathbf{x}_{nse}, \mathbf{s}))$ is obtained.

Temporal Encoder

Table 2 describes the temporal feature extraction network. Weights from the first convolutional layer are shared to extract common temporal signal properties. The output of our temporal extraction process is denoted by: $\mathbf{v}_{eeg} = h_\phi(\widetilde{\mathbf{x}})$ and $\mathbf{v}_{nse} = h_{\phi'}(\mathbf{x}_{nse})$ for encoded eeg and nse features.

Fusion Network

The outputs of the temporal encoder are concatenated channel-wise with state and NSE information $\mathbf{v} = [\mathbf{v}_{eeg}; \mathbf{v}_{nse}; s]$, where the state is broadcast for each temporal feature value. A single layer, fully connected, network $\mathbf{u} = MLP(\mathbf{v})$ is used to fuse channel information over time for each time step. The output $\mathbf{u}$ is passed through our multimodal fusion encoding network (Table 3). Note that we can adjust, add, or remove other modalities of input by modifying the representations concatenated to $\mathbf{v}$ and fused through $\mathbf{u}$ with ease.

For all evaluations 64 dimensions are used as the latent representation output by the last layer of the fusion network along the temporal dimension. For datasets with a smaller number of temporal time steps, the $min(64, T/1.6)$ is used.

Decoder Network

The network is parameterized by $\mathbf{x}' = g_{\theta'}(\mathbf{z})$ using the network given in Table 4. At a very high level it approximates the opposite order of layers presented by the encoder network to obtain the non-corrupt signal. The final output $\mathbf{x}'$ is mapped to the noise-free input via the reconstruction loss:

$$\mathcal{L}_r = -\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{x'}_i\|_1^2. \tag{1}$$

Classification Network

The classification network is a single layer fully connected network which maps the latent vector z to a softmax distribution of 3 classes $\mathbf{y}' = softmax(\ MLP_\rho(z))$. The negative log likelihood is utilized to maximize the correct class distribution:

$$\mathcal{L}_c = -\mathbb{E}[\log p(\mathbf{y}|\mathbf{x})] \tag{2}$$

Joint Loss Function

The final optimization function is a linear combination of the reconstruction and classification loss:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_r, \tag{3}$$

Where λ is an adjustable weight that can be annealed. For all experiments, we set λ to be linearly annealed from 1e−2 to 1e−5 over 5 training epochs. While other functions for annealing are possible, we did not evaluate them for this study.

5 EXPERIMENT SETUP

Two methods of **preprocessing** were used to examine the performance as well as noise tolerance properties of our

model. For each of 1Hz-Bike and 2Hz-Bike data (the two conditions of data collection previously described), the mastoid electrodes were used as the reference electrodes. The original data were also band-pass filtered between .1Hz-255Hz. We analyze a "denoised" version of the dataset, 1Hz-Bike-Filtered and 2Hz-Bike-Filtered, by applying a band-pass filter from 1-40Hz and down sampling to 128Hz. This step removes the high (40Hz+) and low (.1Hz-1Hz) frequency noise as the P300 ERP signal is known, a priori, to be within the 1Hz-40Hz frequency band. Comparing these different preprocessing methods allows us to examine algorithmic behavior under different conditions of noise and prior information.

To evaluate the accuracy and robustness of our algorithm, we split the data in two ways. For **subject-independent** splitting, the post-processed data is split into 80% training and 20% testing instances. We do this for each user stratifying by class distribution. The training data for each user is concatenated and shuffled to create a large training dataset. The same is done for the test dataset. Model and parameter tuning is conducted by randomly splitting 10% of the data from the training set for validation. This method for data splitting was used because this is common practice for current machine learning methodologies, as well as its more challenging condition over single-subject within-subject classification.

Cross-subject splitting allows us to analyze the generalizability of a technique to novel users. We follow the procedure from [28] for subject splitting. Due to our smaller subject pool, we choose 1 subject iteratively and select an additional subject randomly. The remaining 10 subjects are used for training. This process is repeated 12 times so that each subject's test data is tested at least once. We set the training epochs to be 150 when validation accuracy appears to have converged for all models.

Algorithm Comparison

The MCANN model is compared against a traditional non-deep learning approach as well as a previous state-of-the-art deep learning model for ERP classification. For the traditional approach, xDAWN with 5 spatial filters was trained on the EEG data for each class, estimate covariance matrices, and project them into tangent space. Classification is performed using logistic regression with Riemannian distance [5]. This is similar to the technique used to win the Kaggle BCI challenge.

For the deep-learning model, MCANN is compared against EEGNet [28], a CNN architecture which performs comparably to state-of-the-art methods on a number of BCI tasks. For a fair comparison and to study the effects of multimodal signals on existing architectures, two versions of EEGNet are used. The EEGNet (UM) is a unimodal model which is only trained on EEG data. EEGNet (MM) is a multimodal model where we concatenate state information and non-scalp electrodes to the input.

Training and Setup

Training was conducted using a dropout of .25, the adam optimizer with an L2 weight decay of 1e−8, and a learning rate of 1e−3. All hyperparameters were tuned on the validation set of subject independent splitting and kept same throughout evaluation. Early stopping was used during tuning.

Training and evaluation was conducted on a single AMD 2700X with a single NVidia RTX 2070. We measure the performance of running a classification on the test set with a mini-batch size of 1 to simulate how samples would be received during a real-time scenario. Running a single end-to-end evaluation of a single sample takes 39 ms.

6 RESULTS AND DISCUSSION

We examine the macro-averaged precision, recall and F1-scores of all algorithms. Table 5 shows the classification results averaged across all physical activity conditions. We compare all algorithms using both methods of pre-processing for two different data collection parameters (1 Hz and 2 Hz).

Subject-Independent Evaluation

While xDAWN+RG provided the best performance in recall for one of the four conditions, the MCANN model exhibited the best performance in F1-score and in all metrics for all other conditions. In this study, our overall F1-score for subject-independent classification improved 6.28 points or approximately 10% in performance.

Figure 4 provides a confusion matrix for the 1Hz-Bike-Filtered condition. Precision and recall values for recognizing the distractor signal improved with MCANN for the high-intensity exercise condition. Additionally, the decrease in performance between resting and high-intensity exercise (higher noise) for target recognition is noted in this and other experiments. For the four cases we analyzed, the true positive rate for distractors showed the greatest increase.

Noise Robustness. We study the effects of noise on algorithms by changing the sampling and filtering parameters. Our dataset contains noise both within the known ERP frequency (1-40hz) and outside. Future applications of algorithms to EEG might make use of these additional data ranges, potentially preventing the bulk filtering of large frequency bands.

All algorithms (with the exception of the 1 Hz case for xDAWN+RG) typically demonstrated approximately 3% increase in performance metrics over unfiltered data (Table 5). However, when given noisy data, MCANN scores higher on the overall F1-score than other methods under filtered conditions. This suggests MCANN has high tolerance to noise.

Table 5: Summary results for all conditions under subject independent splitting. Average percentage metrics for (R)ecall, (P)recision, and (F1) score reported. Bold signifies best performance.

	1Hz-Bike			2Hz-Bike			1Hz-Bike-Filtered			2Hz-Bike-Filtered		
Method	R	P	F1	R	P	F1	R	P	F1	R	P	F1
xDAWN+RG	**74.16**	54.40	62.76	64.7	46.93	54.40	70.52	52.32	60.07	67.59	49.33	57.03
EEGNet (UM)	64.98	55.99	60.16	58.82	48.63	53.24	68.18	58.29	62.85	66.94	52.00	58.53
EEGNet (MM)	64.98	57.86	61.21	62.04	50.59	55.73	71.72	57.42	63.78	67.70	54.33	60.28
MCANN (Ours)	69.62	**61.92**	**65.55**	**67.09**	**56.33**	**61.24**	**75.33**	**62.31**	**68.20**	**71.92**	**58.27**	**64.38**

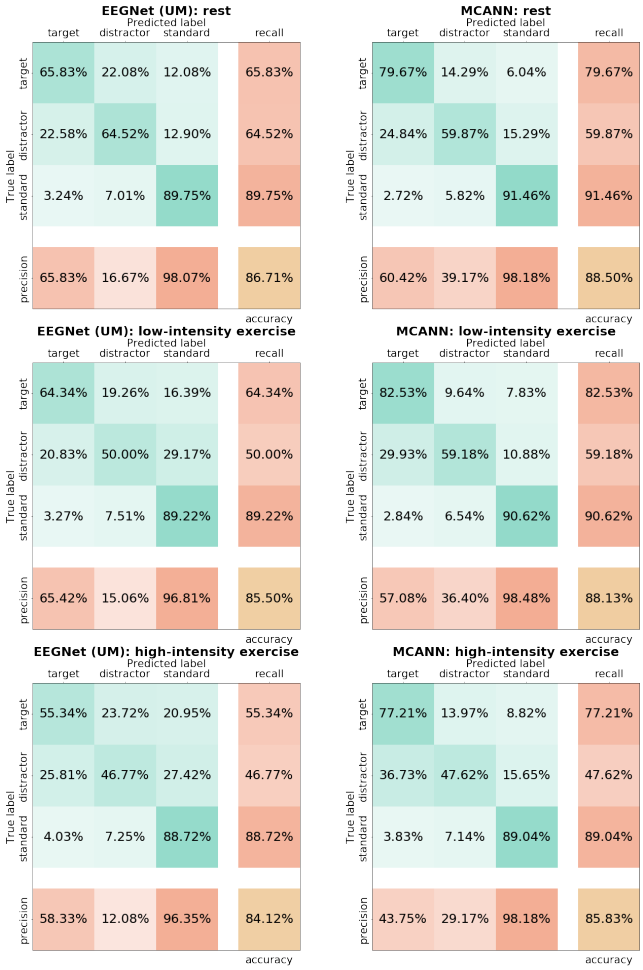

Figure 4: Detailed breakdown of performance for the 1Hz-Bike-Filtered, subject-independent condition.

Effect of Exercise Intensity. All algorithms for all evaluated conditions and preprocessing methods experienced a drop in performance from resting to low- or high- intensity activity. Figure 4 provides an example comparison for the user context versus performance. While MCANN also experiences a drop in performance, it more than doubles the precision for prediction of the distractor in the 1Hz-Bike-Filtered scenario.

Table 6 examines the algorithmic performance over the three conditions. We see that our model and EEGNet (MM) demonstrate an improvement in performance, especially in the high-intensity exercise (higher noise) condition. Our algorithms produced a greater increase in performance under noisy conditions over previous state-of-the-art classifiers, indicating an improved tolerance to noise. EEGNet (MM) also showed improvements over it's UM variant.

Effect of Multimodality. There is some evidence to suggest that the addition of multimodal information can improve classification performance. When comparing the EEGNet (UM) model to EEGNet (MM) performance, we see on average a 1.5 point improvement in F1 score in Table 5. Our proposed method, which also uses the additional modalities performs best on precision and recall scores and leads to an average 6 point improvement in F1 over EEGNet (UM).

Looking at the confusion matrix in Figure 4, we see that EEGNet (UM) target prediction true positives drops by 10 percentage points between the resting and high-intensity exercise conditions. However, for our multimodal approach, we maintain reasonable performance for target predictions and only drop by about 2 points. This suggests that the extra modalities may enhance target signal detection.

Cross-Subject Evaluation

Cross subject evaluation is conducted on the best overall algorithm performance case (1Hz-Bike-Filtered) and the worst overall algorithmic performance case (2 Hz-Bike) from subject-independent evaluation.

A two-sample t-test assuming equal variances is used. We report p-values and effect size using Cohen's d. Our method performs significantly better on metrics against xDAWN+RG, the traditional approach, with greater accuracy (p=0.005, d=1.28), precision (p=0.008, d=1.19), recall (p=0.031, d=0.94), and F1-score (p=0.006, d=1.25).

When compared to EEGNet (UM), our model performs better on accuracy (p=0.014, d=1.09), recall (p=0.002, d=1.44), and F1-score (p=0.032, d=0.93). We did not find any significant difference for precision. Likewise, when compared to EEGNet (MM), we perform significantly better on accuracy (p=0.027, d=0.97), recall (p=0.005, d=1.26), and F1-score

Table 6: 2Hz-Bike, subject independent condition with percentage difference from xDAWN+RG.

User Context	Resting	Low	High
Precision			
xDAWN+RG	52.65	43.7	44.46
EEGNet (UM)	49.89 (-5%)	49.44 (+13%)	46.55 (+5%)
EEGNet (MM)	52.08 (-1%)	49.22 (+13%)	50.47 (+14%)
MCANN (Ours)	**59.93** (+14%)	**54.75** (+25%)	**54.34**(+22%)
Recall			
xDAWN+RG	**70.67**	60.75	59.43
EEGNet (UM)	64.73 (-8%)	60.35(-1%)	53.12 (-11%)
EEGNet (MM)	68.21 (-3%)	59.62(-2%)	59.54 (+0%)
MCANN (Ours)	69.8 (-1%)	**67.8** (+12%)	**63.38** (+7%)

Table 7: Cross subject evaluation on the 1Hz-Bike-Filtered condition. ± standard error is reported.

Method	Resting	Low	High
Precision			
xDAWN+RG	49.44 ± 1.34	46.95 ± 1.24	46.08 ± 1.02
EEGNet (UM)	52.01 ± 1.26	51.34 ± 1.16	48.34 ± 1.17
EEGNet (MM)	52.42 ± 1.37	51.81 ± 1.23	48.51 ± 1.29
MCANN (Ours)	**56.01 ± 1.85**	**54.65 ± 1.65**	**49.33 ± 1.56**
Recall			
xDAWN+RG	68.52 ± 1.86	63.71 ± 2.04	59.31 ± 1.96
EEGNet (UM)	68.38 ± 1.42	61.78 ± 1.36	56.18 ± 1.81
EEGNet (MM)	68.35 ± 1.16	62.46 ± 1.56	56.41 ± 1.79
MCANN (Ours)	**72.02 ± .89**	**68.52 ± 1.41**	**62.29 ± 1.41**

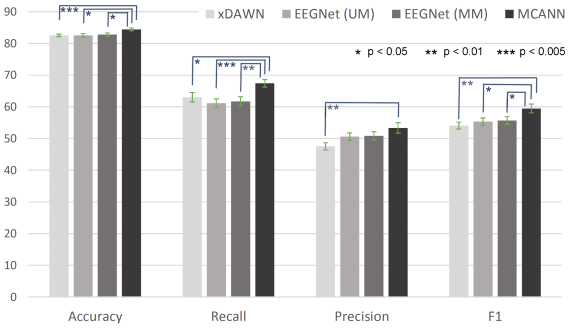

Figure 5: 1Hz-Bike-Filtered performance measures for each algorithm: Accuracy, Recall, Precision, and F1-metric. Error bars show standard error. Brackets indicate p-value significance groups from paired t-tests.

(p=0.05, d=0.85), but there were no significant differences on the precision metric.

Additional significance tests were computed to compare across biking condition. When comparing to EEGNet (UM) we see significant increases for metrics on accuracy (p=0.013, d=1.10), recall (p=0.003, d=1.34), and F1 (p=0.028, d=0.96) for the low condition. For the high activity condition when compared to EEGNet (UM) we perform significantly better on accruacy (p=0.021, d=1.01) and recall (p=0.021, d=1.02). No other significant differences were found when compared to EEGNet (UM). Under resting conditions, no significant difference was found between EEGNet (UM) and our model.

Similar tests are conducted between MCANN and xDAWN+RG. We generally see significantly better performance on almost all metrics for our model. On the high intensity condition, we see accuracy ($p<0.005$, d=.64), precision ($p<0.005$, d=.70), recall ($p<0.005$, d=.49), and F1-Score ($p<0.005$, d=.68). In the low condition, we see improvements in accuracy ($p<0.005$, d=1.6), precision ($p<0.005$, d=1.4), and F1-Score ($p<0.005$, d=1.4). In the resting condition, we found significant improvements in accuracy (p=0.004, d=1.32), precision (p=0.01, d=1.16), and F1-score (p=0.007, d=1.22). Both tests of significance on EEGNet(UM) and xDAWN+RG indicate MCANN's stronger tolerance to noise.

When looking at the worst case scenario with the 2Hz-Bike dataset, we did not find any significant differences among any of the methods. We believe that, due to the minimal amount of processing and higher presentation frequency, the 2Hz-Bike dataset has a very large amount of individual subject variance outside the traditional 1-40Hz range, making it difficult for any model to generalize to new subjects. Some of these differences may be due to differences in individual motion patterns such as riding posture and cadence. These factors introduce unique noise patterns that compound the already challenging task of performing classification across users. These results highlight the need for more robust multimodal sensors to be used in conjunction with EEG sensors. We intend to investigate these inter-user differences in future research.

7 CONCLUSION

In this paper we presented a challenging dataset for developing BCI classification algorithms. We provided a novel method for classifying EEG signals under conditions that varied dramatically with regard to noise. We observed significant improvements in our test set during cross-subject evaluation when compared to previous state-of-the-art techniques. Additionally, our new algorithm is capable of incorporating additional modalities for improved classification of brain data. Future work will include (1) the application of our classifier to online BCI scenarios that involve motion, such as navigation of real-life or virtual environments, and (2) testing classifier performance during other types of physical activity that may involve more extreme head and body movements.

ACKNOWLEDGMENTS

This work was supported in part by NSF awards IIS-1845587 and DGE-1258507, as well as ONR award N00014-16-1-3002. The authors would like to thank Cristopher Garduno, Yimeng Liu, Lu Han, and Jayleen Li for help with data processing and valuable discussions.

REFERENCES

[1] Gregory D Abowd and Elizabeth D Mynatt. 2000. Charting past, present, and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 7, 1 (2000), 29–58.

[2] Aurélien Appriou, Andrzej Cichocki, and Fabien Lotte. 2018. Towards robust neuroadaptive HCI: exploring modern machine learning methods to estimate mental workload from EEG signals. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, LBW615.

[3] P Aricò, G Borghini, G Di Flumeri, N Sciaraffa, and F Babiloni. 2018. Passive BCI beyond the lab: current trends and future directions. *Physiological measurement* 39, 8 (2018), 08TR02.

[4] John Atkinson and Daniel Campos. 2016. Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. *Expert Systems with Applications* 47 (2016), 35–41.

[5] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. 2012. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering* 59, 4 (2012), 920–928.

[6] Ali Bashashati, Mehrdad Fatourechi, Rabab K Ward, and Gary E Birch. 2007. A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural engineering* 4, 2 (2007), R32.

[7] G. Borg. 1970. Perceived exertion as an indicator of somatic stress. *Scandinavian Journal of Rehabilitation Medicine* 2, 2 (1970), 92–98. https://www.scopus.com/inward/record.uri?eid=2-s2.0-0014895192&partnerID=40&md5=fa8e27f6d6bf24a2f47dac587715f66c cited By 2796.

[8] Raymond C Browning, Emily A Baker, Jessica A Herron, and Rodger Kram. 2006. Effects of obesity and sex on the energetic cost and preferred speed of walking. *Journal of applied physiology* 100, 2 (2006), 390–398.

[9] Tom Bullock, Hubert Cecotti, and Barry Giesbrecht. 2015. Multiple stages of information processing are modulated during acute bouts of exercise. *Neuroscience* 307 (2015), 138–150.

[10] Tom Bullock, James C Elliott, John T Serences, and Barry Giesbrecht. 2017. Acute exercise modulates feature-selective responses in human cortex. *Journal of cognitive neuroscience* 29, 4 (2017), 605–618.

[11] Hubert Cecotti, Miguel P Eckstein, and Barry Giesbrecht. 2014. Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering. *IEEE transactions on neural networks and learning systems* 25, 11 (2014), 2030–2042.

[12] Hubert Cecotti and Axel Graeser. 2008. Convolutional neural network with embedded Fourier transform for EEG classification. In *2008 19th International Conference on Pattern Recognition*. IEEE, 1–4.

[13] Hubert Cecotti and Axel Graser. 2011. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE transactions on pattern analysis and machine intelligence* 33, 3 (2011), 433–445.

[14] H. Cecotti, A. Marathe, and A. Ries. 2015. Optimization of single-trial detection of event-related potentials through artificial trials. *IEEE Trans. on Biomedical Engineering* 62, 9 (2015), 2170–6.

[15] Chih-Ming Chen, Jung-Ying Wang, and Chih-Ming Yu. 2017. Assessing the attention levels of students by using a novel attention aware system based on brainwave signals. *British Journal of Educational Technology* 48, 2 (2017), 348–369.

[16] Guy Cheron, Géraldine Petit, Julian Cheron, Axelle Leroy, Anita Cebolla, Carlos Cevallos, Mathieu Petieau, Thomas Hoellinger, David Zarka, Anne-Marie Clarinval, et al. 2016. Brain oscillations in sport: toward EEG biomarkers of performance. *Frontiers in psychology* 7 (2016), 246.

[17] Luis F Ciria, Pandelis Perakakis, Antonio Luque-Casado, and Daniel Sanabria. 2018. Physical exercise increases overall brain oscillatory activity but does not influence inhibitory control in young adults. *Neuroimage* 181 (2018), 203–210.

[18] Andy Cockburn, Carl Gutwin, Joey Scarr, and Sylvain Malacria. 2015. Supporting novice to expert transitions in user interfaces. *ACM Computing Surveys (CSUR)* 47, 2 (2015), 31.

[19] Gerhard Fischer. 2001. User modeling in human–computer interaction. *User modeling and user-adapted interaction* 11, 1-2 (2001), 65–86.

[20] Nir Friedman, Tomer Fekete, Ya'akov Kobi Gal, and Oren Shriki. 2019. EEG-based Prediction of Cognitive Load in Intelligence Tests. *Frontiers in Human Neuroscience* 13 (2019), 191.

[21] Fabien Grego, Jean-Marc Vallier, Maya Collardeau, Stéphane Bermon, Patricia Ferrari, Mirande Candito, Pascale Bayer, Marie-Noëlle Magnié, and Jeanick Brisswalter. 2004. Effects of long duration exercise on cognitive function, blood glucose, and counterregulatory hormones in male cyclists. *Neuroscience letters* 364, 2 (2004), 76–80.

[22] Joseph T Gwin, Klaus Gramann, Scott Makeig, and Daniel P Ferris. 2010. Removal of movement artifact from high-density EEG recorded during walking and running. *Journal of neurophysiology* 103, 6 (2010), 3526–3534.

[23] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation* 18, 7 (2006), 1527–1554.

[24] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[25] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[26] Laurens R Krol and Thorsten O Zander. 2017. Passive BCI-based Neuroadaptive Systems.. In *GBCIC*.

[27] Naveen Kumar and Jyoti Kumar. 2016. Measurement of cognitive load in HCI systems using EEG power spectrum: an experimental study. *Procedia Computer Science* 84 (2016), 70–78.

[28] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering* 15, 5 (2018), 056013.

[29] Lun-De Liao, Chi-Yu Chen, I-Jan Wang, Sheng-Fu Chen, Shih-Yu Li, Bo-Wei Chen, Jyh-Yeong Chang, and Chin-Teng Lin. 2012. Gaming control using a wearable and wireless EEG-based brain-computer interface device with novel dry foam-based sensors. *Journal of neuroengineering and rehabilitation* 9, 1 (2012), 5.

[30] Fu-Ren Lin and Chien-Min Kao. 2018. Mental effort detection using EEG data in E-learning contexts. *Computers & Education* 122 (2018), 63–79.

[31] Ning-Han Liu, Cheng-Yu Chiang, and Hsuan-Chin Chu. 2013. Recognizing the degree of human attention using EEG signals from mobile sensors. *Sensors* 13, 8 (2013), 10273–10286.

[32] Monika Lohani, Brennan R Payne, and David L Strayer. 2019. A review of psychophysiological measures to assess cognitive states in real-world driving. *Frontiers in human neuroscience* 13 (2019).

[33] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. 2018. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering* 15, 3 (2018), 031005.

[34] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. 2007. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of neural engineering* 4, 2 (2007), R1.

[35] Perrin Margaux, Maby Emmanuel, Daligault Sébastien, Bertrand Olivier, and Mattout Jérémie. 2012. Objective and Subjective Evaluation of Online Error Correction during P300-Based Spelling. *Advances in Human-Computer Interaction* 2012 (dec 2012), 1–13. https://doi.org/10.1155/2012/578295

[36] Kevin Nathan and Jose L Contreras-Vidal. 2016. Negligible motion artifacts in scalp electroencephalography (EEG) during treadmill walking. *Frontiers in human neuroscience* 9 (2016), 708.

[37] John Polich. 2007. Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology* 118, 10 (2007), 2128–2148.

[38] John Polich and JoséR. Criado. 2006. Neuropsychology and neuropharmacology of P3a and P3b. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology* 60 2 (2006), 172–85.

[39] H. Raza, S.M. Rathee, D.and Zhou, H. Cecotti, and G. Prasad. 2019. Covariate Shift Estimation based Adaptive Ensemble Learning for Handling Non-Stationarity in Motor Imagery related EEG-based Brain-Computer Interface. *Neurocomputing* 343 (2019), 154–166.

[40] Pedro Reis, Felix Kluge, Florian Gabsteiger, Vinzenz Tscharner, and Matthias Lochmann. 2014. Methodological aspects of EEG and Body dynamics measurements during motion. *Frontiers in human neuroscience* 8 (03 2014), 156. https://doi.org/10.3389/fnhum.2014.00156

[41] Matthias Rötting, Thorsten Zander, Sandra Trösterer, and Jeronimo Dzaack. 2009. Implicit interaction in multimodal human-machine systems. In *Industrial Engineering and Ergonomics*. Springer, 523–536.

[42] Ahmed Ben Said, Amr Mohamed, Tarek Elfouly, Khaled Harras, and Z Jane Wang. 2017. Multimodal deep learning approach for joint EEG-EMG data compression and classification. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 1–6.

[43] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping* 38, 11 (2017), 5391–5420.

[44] Andrew B Schwartz, X Tracy Cui, Douglas J Weber, and Daniel W Moran. 2006. Brain-controlled interfaces: movement restoration with neural prosthetics. *Neuron* 52, 1 (2006), 205–220.

[45] Sergei L Shishkin, Yuri O Nuzhdin, Evgeny P Svirin, Alexander G Trofimov, Anastasia A Fedorova, Bogdan L Kozyrskiy, and Boris M Velichkovsky. 2016. EEG negativity in fixations used for gaze-based control: Toward converting intentions into actions with an eye-brain-computer interface. *Frontiers in neuroscience* 10 (2016), 528.

[46] P. Simard, D. Steinkraus, and J.C. Platt. 2003. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. (Aug. 2003), 958–962.

[47] Michael Tangermann, Klaus-Robert MÃijller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai Miller, Gernot Mueller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois SchlÃűgl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. 2012. Review of the BCI Competition IV. *Frontiers in Neuroscience* 6 (2012), 55. https://doi.org/10.3389/fnins.2012.00055

[48] Jose Antonio Urigüen and Begoña Garcia-Zapirain. 2015. EEG artifact removalâĂŤstate-of-the-art and guidelines. *Journal of neural engineering* 12, 3 (2015), 031001.

[49] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. 2002. Brain–computer interfaces for communication and control. *Clinical neurophysiology* 113, 6 (2002), 767–791.

[50] Yasuo Yagi, Kerry L Coburn, Kristi M Estes, and James E Arruda. 1999. Effects of aerobic exercise and gender on visual and auditory P300, reaction time, and accuracy. *European journal of applied physiology and occupational physiology* 80, 5 (1999), 402–408.

[51] Zhong Yin and Jianhua Zhang. 2017. Cross-session classification of mental workload levels using EEG and an adaptive deep learning model. *Biomedical Signal Processing and Control* 33 (2017), 30–47.

[52] Thorsten O Zander, Laurens R Krol, Niels P Birbaumer, and Klaus Gramann. 2016. Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity. *Proceedings of the National Academy of Sciences* 113, 52 (2016), 14898–14903.