An Open, Diverse and Machine Learning Ready P300-based Brain-Computer Interface Dataset

Xinlin J. Chen[1] **Riki Shimizu**[1] **Chance Fleeting**[1] **Thomas Balmat**[2] **Eric Jiayuan Qi**[1]
William Y. Luqiu[1] **Wei Wu**[1] **Leslie M. Collins**[1] **Eric W. Sellers**[3] **Boyla O. Mainsah**[1]
[1]Department of Electrical and Computer Engineering, Duke University, USA
[2]Office of Information Technology, Duke University, USA
[3]East Tennessee State University, USA

Abstract

The P300-based brain-computer interface (BCI) is one of the most commonly researched BCIs for communication. We present an open P300-based BCI speller dataset of 170 participants that provides data in an enriched and standardized format, with study metadata, demographic information when available and BCI data levels that align with developing IEEE BCI data standards to facilitate reusability. Our BCI dataset is a collection of data curated from our previous P300 speller studies from over 15years of BCI research, encompassing a wide range of user performance levels and experimental conditions, including longitudinal data for simulating extended BCI use. To address the current under-representation of target end users in BCI data, our dataset includes data from 39 participants with amyotrophic lateral sclerosis (ALS), a target BCI user population. We demonstrate the broad utility of our dataset with experiments of machine learning tasks for character selection and error detection in P300 BCI spellers.

1 Introduction

Brain-computer interfaces (BCIs) record and process signals associated with brain activity from users, translating neurophysiological data into commands to control external devices (1). There exist a multitude of BCI applications, such as BCIs for communication in people with severe neuromuscular limitations, or for restoration of voluntary movement in individuals with paralysis. However, current noninvasive BCIs have relatively low communication rates, as they rely on processing and interpreting noisy neurophysiological data. Moreover, the neural signal components from which relevant information is extracted to effectuate BCI control are highly variable. Presently, improving BCI communication efficiency is an area of significant research interest.

A key component of the development process for BCI algorithms involves performing simulations with data previously collected from other BCI users (1). Such simulations enable the evaluation of a wide array of BCI algorithms under consideration prior to, or even in place of, online testing, to identify a few promising candidates for subsequent online testing. However, a core challenge in BCI research is the relatively low-resource environment and time consuming nature of BCI data acquisition, especially when compared to publicly available datasets for popular machine learning applications such as computer vision and natural language processing. Rather than collecting data in-house, most BCI research rely on publicly available datasets for analyses.

Need for Big Data in BCI Research. Due to high data variability across individuals, the standard approach for BCI machine learning is user-specific. Alternatively, data collected from other BCI users can be leveraged for training generalized models, pretraining models to facilitate transfer learning or analyzed in aggregate to identify performance trends or users with similar profiles. Given the

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

successes in other fields, recent years have seen an increased interest in using deep learning for BCI applications (2; 3; 4); such models benefit from learning from a large and diverse user dataset (5), especially for generalizing to novel BCI users. However, most BCI datasets typically have a small number of participants and lack multi-session data for longitudinal assessments ((6), Table 1).

Need for BCI Datasets with Target BCI End Users. While most online BCI studies are conducted with abled-bodied individuals for practical reasons, the gold standard for evaluating BCI algorithms is online studies with target end users, such as individuals with amyotrophic lateral sclerosis (ALS). However, only a few BCI research groups, including our group, conduct BCI studies in target BCI end user populations (7; 8) and these data are rarely disseminated. It is vital to perform simulations with data obtained from target BCI end users to better reflect technical specifications, data conditions and performance trends in this population. Thus, there is a need to address the severe under-representation of target end users in current BCI datasets, which are highly biased with able-bodied individuals.

Need for a Standard BCI Data Format. A potential solution to the limited BCI data problem is merging BCI datasets across various repositories. However, this is challenging due the lack of well-defined BCI data definitions, lack of or limited data documentation and differences in file format and BCI data content across BCI datasets (6). Most BCI research groups use the open-source BCI2000 software platform (9; 10), which is supported by the National Institutes of Health (NIH). Data recorded in BCI2000 are stored in a native *.dat* file format, which is readable using a proprietary MATLAB package provided by the BCI2000 developers; consequently, BCI2000 data files are typically reformatted and shared in other file formats with adhoc BCI data dictionaries. Important data attributes related to biosignal acquisition, such as sensor technology and participant demographics, are usually missing from BCI data files or provided in a separate file, creating issues with portability (6). The IEEE P2731 working group is currently developing new standards for a unified BCI terminology (11), including for BCI data storage and sharing (6).

To address the gaps in current BCI research data, we have developed a highly structured BCI dataset based on the currently developing standards for data storage and sharing from the IEEE P2731 Working Group on BCIs (6). Our BCI research focuses on the P300 speller, a widely researched BCI for communication. We have a large collection of propriety BCI2000 data files generated from several online P300 speller studies from over 15 years of BCI research funded by the National Institutes of Health (NIH). We have curated, cleaned and engineered the BCI data to carefully design a P300 BCI dataset and our key contributions in this work are summarized as follows:

1. We present an open, large, diverse P300 BCI dataset of 170 participants, including 39 participants with ALS. Our BCI dataset encompassing a wide range of user performance levels and experimental conditions during P300 speller use. The final BCI dataset will be made publicly available and will be unique in its size, coherency, diversity, and the inclusion of high number of target BCI end users.
2. We provide a new BCI data dictionary for the P300 speller that includes all BCI data levels to align with currently developing IEEE P2731 WG standards for BCI data storage and sharing (6). When available, participant-specific demographic information (age, sex, race, ethinicity, ALS diagnosis, ALS severity score) is embedded in the BCI data file.
3. We present results to demonstrate the wide range in user performance and the reusability of our BCI dataset in experiments with machine learning tasks. We investigate the effect of data preprocessing on model performance for character selection and the feasibility of a generalized model for error detection in P300 spellers.

2 Related Work

We limited our literature search to publicly available visual P300 speller datasets with electroencephalography (EEG) data for relevance to our BCI dataset; Table 1 summarizes the characteristics of visual P300 speller BCI datasets with EEG data (12; 13; 14; 15; 16; 17; 18; 19; 20). Only one study included participants with ALS (19). Most datasets have between 1 to 13 participants, except (20) with 55 participants. All but one of the datasets provide data files in binary MATLAB format. (15) provides files in the open European data format (EDF); while we also provide EDF files in our dataset, our BCI data dictionary empassess all BCI data levels recommended by the IEEE P2731 WG on BCI standards and includes demographic information in the data file when available. Most of the

Table 1: Comparison of publicly available visual P300 BCI speller datasets and our dataset.

Dataset	N	ALS Population	No. of Sessions	Stimulus Paradigm(s)	File Format	User-specific Demographics
Blankertz et al., 2004 (12)	1	No	3	RC	MAT	-
Blankertz et al., 2006 (13)	2	No	5	RC	MAT	-
Guger et al., 2009(14)	10	No	1	RC, SC	MAT	-
Citi et al., 2010 (15)	12	No	1	RC	EDF	-
Treder et al., 2011 (16)	13	No	1	Regional	MAT	-
Aloise et al., 2012 (17)	10	No	1	Center, RC	MAT	{Age, Sex}, In a separate file
Acqualagna et al., 2013 (18)	12	No	1	RSVP	MAT	-
Won et al., 2022 (20)	55	No	1	RC	MAT	-
Riccio et al., 2013 (19)	8	Yes	1	RC	MAT	{Age, Sex, ALSFRS-r, Onset}, Embedded
Our dataset†	170	Yes	1, 2-5	RC, CB, RD, PB, AD	EDF+	{Age, Sex, Race/Ethnicity, ALSFRS-r}, Embedded

Abbreviations: AD, Adaptive; CB, Checkerboard; EDF, European Data Format; MAT, binary MATLAB files; MS, Multi-session; N, Number for study participants; PB, Performance-Based; RD, Random; RC, Row-Column; RSVP, rapid serial visual presentation; SS, Single Session.

†This is the size of our current dataset. Total size after finalizing dataset development is 300+ participants.

datasets do not provide user-specific demographics; for some studies, summary statistics are available in a related publication (14; 18; 16; 20). One study includes user-specific demographics in a separate file (17). Similar to our dataset, (19) embeds participant-specific demographics in the BCI data file.

Compared to similar public datasets, our proposed P300 speller dataset is unique in having the following attributes: a) open files that provide data in an enriched and IEEE-standardized format to facilitate reusability; b) a sufficiently large number of participants (N = 170) in a single dataset to supply "big data" for machine learning; c) Diverse stimulus presentation paradigms to better model the impact of psychophysical effects of stimuli, such as refractory and distractor effects, during P300 speller simulations; d) the inclusion of data from target BCI end users, namely individuals with ALS (N = 39), which is crucial in addressing the bias in BCI datasets that predominantly contain data from abled-bodied individuals; e) longitudinal data for assessment of BCI algorithms over long-term use; our BCI dataset includes two longitudinal studies with individuals with ALS.

3 Dataset Description

3.1 Ethics and Privacy

All BCI studies were performed under protocols approved by Institutional Review Boards at Duke University, the Duke University Health System and East Tennessee State University. The dataset is anonymized to protect privacy: all personal identifiable information have been removed, each participant has been assigned an identification number and all dates (e.g., data collection, birth date, etc.) have been time-shifted. Participants were made aware on the consent forms that their anonymized data will be publicly shared. Participants with ALS who have a limited means of communication may struggle to convey discomfort or wish to terminate a BCI experiment, which may be risky; we ensured that a familiar person was present during experiments to facilitate communication.

3.2 Data Acquisition

3.2.1 Participants

All participants gave informed consent, either by themselves or via a legally authorized representative (for some participants with ALS), prior to data collection. Participants were compensated for their time with either course credit, cash or gift card payments ($12 to $25 per hour). When available, participant demographics include self-reported age, sex, race, and ethnicity, as well as ALS diagnosis and a revised ALS Functional Rating Scale (ALSFRS-r) score obtained from their medical records. The ALSFRS-r is an instrument for evaluating the degree of functional impairment in individuals with ALS (21) with a range of 0 to 48; a lower value indicates a higher degree of functional impairment.

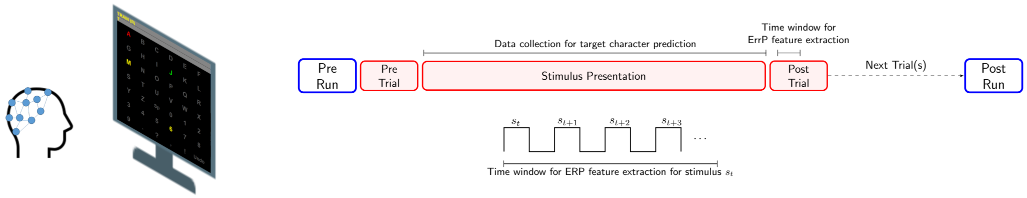
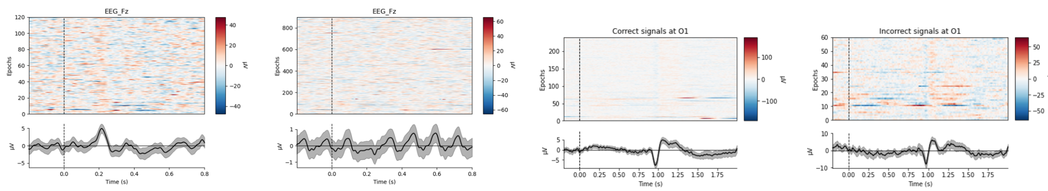


Figure 1: Schematic illustration of a P300 speller and key events during use. To spell a character, the user attends to the target character while stimuli, subsets of characters are illuminated on the screen, are presented as the BCI measures the user's EEG signals.



(a) Target stimuli (b) Nontarget stimuli (c) Correct feedback (d) Incorrect feedback

Figure 2: Example neural responses from a BCI user elicited during P300 speller use after presentation of (a) target and (b) non-target stimuli for target character estimation, and after rpesentation of (c) correct and (b) incorrect BCI feedback. Placeholder figures

3.2.2 Technical Setup

Data were recorded using BCI2000 (9). For participants without ALS, a 9×8 (number of rows $\times$ number of columns) or 6×6 user interface, stimulus duration of 62.5 ms and inter-stimulus interval (ISI) of 62.5 ms were used. For participants with ALS, a 6×6 user interface, stimulus duration of 125ms and ISI of 125 ms were used. Different stimulus presentation rates are typically used for individuals with and without neuromuscular limitations due to psychophysical factors and physical impairments that may impact BCI use (22; 23); for example, individuals with neuromuscular limitations are less tolerant of high stimulus presentation rates and find it easier to navigate a less complex user interface. EEG signals were collected at 256 Hz using gel-based passive electrodes with a 10-20 electrode montage connected to either one or two 16-channel gUSBamp biosignal amplifiers (g.tec medical engineering GmbH). Ground and reference electrodes during EEG signal recordings were placed at the left and right mastoids, respectively. An electrode impedance check was conducted to ensure low impedance (generally < 40 kΩ). Raw EEG data were bandpass filtered (0.5 - 30 Hz) and when applicable, notch filtered (60 Hz) to remove electrical noise. For hybrid BCI use, eye gaze position, eye position, eye distance from the screen and pupil diameter were collected using a Tobii Pro X2-30 (Tobii AB) infrared eye tracker. The eye tracker was calibrated for each participant prior to BCI use. Raw eye tracker data were preprocessed based on the technical specifications in (24).

3.2.3 Experiment Protocols

The current batch of our BCI dataset includes data from 12 P300 speller studies with 170 participants, with 39 participants with ALS. Study procedures are detailed in related publications and a summary is provided in Table 2. In general, all participants performed copy-spelling of predefined tokens (words or number sequence) using the P300 speller; the set of tokens was randomly drawn for each participant. Figure 1 shows a schematic of the experiment setup and key events during BCI use. A user is presented with a set of choices on a speller interface. To select a character, the user focuses on that target character as subsets of characters are illuminated on the screen. The illumination of a subset of characters on the screen represents a visual stimulus event. It is assumed that rarely occurring target stimuli elicits event related potentials (ERPs) with a positive peak around 300 ms (P300) from stimulus presentation. Figure 2a-2b shows example EEG signals following presentation of non-target and non-target stimuli, respectively. The BCI infers the user's intended character by: i) processing EEG data following each stimulus presentation; ii) using a classifier to detect P300 ERPs embedded in EEG data that are elicited in response to target stimuli; and iii) estimating the user's

Table 2: Summary of our previous P300 speller studies included in our dataset.

Study	**No. of Channels**	**Grid‡ Size**	**Stimulus Paradigm(s)**	**No. of Participants**	**ALS Study?**	**Total No. of Character Trials** Train/Test	**No. of Stimuli per Character** Train/Test	**No. of Sessions**
A (25)	32	9×8	RC, CB,RD	13	No	105/105	120/120	1
B (26)	16	6×6	CB	18	Yes	Var/Var	180/180	Var
C (27)	32	9×8	CB	19	No	20/300	120/24	1
D (28)	32	9×8	RC	17	No	36/72	170/Var	1
E*	16	9×8	CB	8	No	54/108	120/Var	1
F (29)	16	6×6	CB	10	Yes	36/Var	120/Var	3
G (30)	16	9×8	CB	20	No	36/Var	120/Var	1
H (31)	16	9×8	CB	16	No	180/120	168/168	1
I†	16	9×8	CB, PB	13	No	48/72	120/Var	1
J (32)	16	6×6	RC, PB	20	No	60/60	72/Var	1
K (33)	16	9×8	CB, AD	5	No	30/60	120/Var	1
L (34)	16	6×6	RC, CB, CB$_c$	11	Yes	90/90	145/145	1

Abbreviations: AD, Adaptive; ALS, amyotrophic lateral sclerosis; CB, Checkerboard; CB$_c$, Checkerboard Color; No., number; PB, Performance-Based; RD, Random; RC, Row-Column; Var, variable.
‡Grid size is specified as number of rows $\times$ number of columns in a matrix layout.
*The experiment protocol of study E is similar to that of study F. †The study protocol of study I is similar to that of study J.

intended character by matching the character presentation patterns to the detected P300 ERPs with a character decoding algorithm.

A P300 speller experiment session consists of a calibration phase and a test phase. During the calibration phase, participants perform copy-spelling with no classifier use and no BCI feedback presented to collect labeled EEG data to train a P300 classifier. During the test phase, the trained P300 classifier is applied and participants perform copy-spelling with the BCI prediction of the target character presented as feedback at after data collection for a character trial. BCI actions that are perceived by the BCI user, such as BCI feedback presentation, are also assumed to modulate the user's brain signals (6) can impact the user's psychological state. Error-related potentials (ErrPs) are neural signal components that are elicited when a person perceives erroneous actions or behavior (35). Thus it is assumed that an ErrP is elicited when the user observes a mismatch between the intended target character and the presented BCI feedback.

3.3 Data Content

We extracted biosignal data and technical parameters from the source BCI 2000 *.dat* files, enriched with additional data elements, such as participant demographics (when available) and study information, and reformatted to the EDF+(36). EDF+ is an open format for multi-channel signals with file specifications that describe technical and personal attributes. EDF+ file readers are available for various software platforms, including MATLAB, R, Python and C++ (37).

We have defined a new data dictionary for the file header and data records in the EDF+ file that encompasses all the BCI data levels outlined in the IEEE P2731 *Standard for a Unified Terminology for Brain-Computer Interfaces* for data storage and sharing (6):

***BCI data level 0*:** This includes information related to signal acquisition, such as brain signals, other biosignals and demographics. All data files contain EEG signals. When available, eye tracker signals (for hybrid BCIs), demographics (age, sex, race, ethnicity, with categories as defined by the NIH (38)), ALS diagnosis and ALSFRS-r score are provided.

***BCI data level 1*:** This includes information about the BCI paradigm that is needed to train the machine learning model. The target character, stimulus type (target or non-target) and stimulus presentation schedule are included for each spelling trial.

***BCI data level 2*:** This includes information related to the BCI prediction. We distinguish between the BCI prediction and the presented BCI feedback because the latter can be different in certain experimental paradigms, such as the fake feedback paradigm described later in this paper.

The EDF+ file header contains a *patient* identification (ID) field with sub-fields to describe individual attributes and a *recording* ID field with sub-fields to include information about the recording setting. Our BCI dataset does not contain personal identifiers. We modified the patient ID sub-fields to include additional demographic information (race/ethinicity, ALS diagnosis, ALS severity) beyond the default options (sex, age) that are inherent to the *.edf* file. We also modified the record ID sub-fields to include the study identifier and experiment session number. Our BCI data dictionary for the *.edf* file based on the proposed IEEE P2731 data levels are provided in Table A1 (Appendix A).

A unique aspect of our BCI dataset is hhe inclusion of several BCI data attributes that facilitate reusability (39). All P300 BCI datasets contain BCI data level 1 elements to train the P300 classifier and to predict the user's target character. The class-specific BCI stimulus distribution across participants is provided in Appendix C.1. Even though the data are generated during the test phase of BCI experiments, none of the P300 BCI datasets from our literature review (Table 1) include BCI data level 2 elements to facilitate ErrP analysis. Our BCI dataset also includes BCI data level 2 elements that are needed for ErrP analysis with binary labels of agreement between the user's intent and the presented BCI feedback. The distribution of the online participant accuracies and stopping times from real-time BCI use are provided in Appendix C.2. The next two sections highlight the *multi-purpose* use of our BCI dataset for P300 ERP and ErrP analyses.

4 Data Analysis

4.1 Signal preprocessing

For each channel, a time window of EEG data was extracted from the task-relevant onset time, which is either BCI stimulus presentation or BCI feedback presentation (see Figure 1). Due to the high dimensionality of raw EEG signals, various dimensionality reduction techniques were applied.

Channel selection. We used either the standard 8-channel subset (Fz, Cz, P3, Pz, P4, PO7, PO8, Oz) (40) or a 16-channel subset (F3, Fz, F4, T7, C3, Cz, C4, T8, CP3, CP4, P3, Pz, P4, PO7, PO8, Oz) for a common number of channels across studies (Table 2).

Data resampling. This included block-averaging of non-overlapping time segments within the EEG time window (40) or downsampling with a decimation filter at a specified factor.

Spatial filtering. xDAWN is a technique to enhance evoked potentials by projecting raw EEG data to a channel subspace that increases the signal to signal-to-noise ratio of evoked responses (41; 42).

4.2 Models

Linear discriminant analysis (LDA). This is the baseline state-of-the art traditional machine learning model that was used in the original P300 speller study (43) and its variants are still popularly used (44), including in our BCI studies. Stepwise LDA is the default classifier in BCI2000 (45).

Convolutional neural network (CNN). EEGNet is a compact CNN architecture that was developed for generic classification in EEG-based BCIs (46). To the best of our knowledge, EEGNet is the only deep learning model in the literature that has been validated against stepwise LDA in an online P300 speller study (5). EEGNet consists of a sequence of layers with temporal, depth-wise, and separable convolutional filters; we used the EEGNet-8,2 architecture, where the notation EEGNet-F_1, D indicates the number of temporal ($F1$) and spatial (D) filters. The specifications for the EEGNet architecture can be varied based on the number of channels and time sample points.

Recurrent neural network. (47) developed a two layer cascade of a CNN to capture spatial information and long short-term memory (LSTM) network to capture temporal information. Similar to EEGNet, the CNN-LSTM architecture allows for a variable number of channels and time sample points. For this analysis, we used the small CNN-LSTM variant in (47).

Model training and evaluation were performed in Python on virtual machines with SYS-1029GQ-TNRT central processing units (CPUs) and RTX A5000 graphics processing units (GPUs).

4.3 Tasks

4.3.1 Impact of Signal Preprocessing on BCI Performance of Deep Learning Models

Dimensionality reduction is a standard preprocessing step when using traditional machine learning models for BCI applications. In contrast, the full channel set and raw EEG data with minimal preprocessing are typically used as input features to deep learning models (2; 3; 4); this is based on the rationale of leveraging model complexity to automatically learn robust feature representations from the raw data without the need for manual feature extraction. However, the high EEG data dimensionality has implications on model performance given the typical amounts of available user-specific training data relative to the number of trainable parameters in a BCI classifier model. The number of trainable parameters is typically in the order of 10^2 for traditional machine learning models vs. 10^3 to 10^5 for deep learning models, e.g., (46), while the amount of user-specific data from the training phase ranges from 2 to 6 $\times 10^3$ observations to train the P300 classifier.

We conducted experiments to investigate the impact of various signal preprocessing on the performance of deep learning models. We selected neural networks from the literature that allowed for variable input sizes based on the author specifications to accommodate dimensionality reduction with signal preprocessing. Raw features were extracted from a time window of 0.625 seconds from stimulus onset across all participants. The preprocessing steps applied included: channel subset (8, 16), downsampling decimation factor (none, 4, 8) and spatial filtering (with xDAWN, without xDAWN). We used the LDA model in the *scikit-learn* package, the EEGNet,8-2 model (46) provided by the model developers via GitHub (48), and our implementation of the CNN-LSTM based on the specification in (47), with a modification from a sigmoid to a tangent function for the last activation layer. For both deep learning models, subject-specific P300 classifiers were trained on data from the calibration phase with a cross-validation split of 0.1 for 100 epochs using the Adam optimizer, a binary cross-entropy loss and an initial learning rate of 1e-4.

Data from the test phase were used to predict the target character. The BCI character decoder is the cumulative sum of classifier scores of a character during a trial. To select a new target character, all characters are initially assigned a zero score. After each stimulus presentation, the trained classifier is applied to the corresponding EEG features and only characters that are present in the current stimulus event receive a character score update. After data collection is terminated, the character with the maximum cumulative classifier score is selected as the user's target character estimate.

4.3.2 Towards Generalized Error Detection in BCI Spellers

ErrPs have been proposed as endogenous feedback signals from the BCI user to assess the accuracy of BCI predictions (35). However, there has been limited use of ErrPs within the context of P300-based BCIs due to low accuracy with single trial ErrP detection and the time needed to obtain sufficient training data (35). For example, a typical character trial of a P300-based BCI that yields 120 samples (based on the number of stimuli presented) to train a P300 classifier but only one sample per character (after BCI feedback is presented) to train an ErrP classifier. Moreover, for BCI users with high accuracy, the error class could be rare or nonexistent.

To ensure enough training data and samples of each class, a separate calibration phase is performed for the ErrP classifier: the BCI user performs copy-spelling at a reduced data collection limit and fake BCI feedback at a predefined error rate is presented (35). Study C in our dataset ((27)) is a fake BCI feedback paradigm study: the experiment includes a conventional calibration phase to train a P300 classifier and a fake feedback phase that provides a sufficient amount of data to train an ErrP classifier. However, a separate calibration phase for the ErrP classifier is still cumbersome. As a first step towards a generalized error detection model for the P300 speller, we performed a preliminary analysis to investigate the transferrability of ErrP classifiers trained on data from other BCI users and applied to novel users without retraining. This included two approaches: using an ensemble of user-specific classifiers or a generic classifier trained on pooled data from other BCI users.

ErrP classifier models were trained on data from the fake feedback phase from participants in study C (*source* users) and applied to data from other subjects (*novel* users) in the same study or other studies. Data from the fake feedback phase consists of 300 character observations with an error rate of 20%. First, we investigated the performance of ErrP classifiers trained on various feature sets. We used the LDA model from the *scikit-learn* package and our implementation of the EEGNet-8,2 model in PyTorch based on (46). To generate the ErrP classifier ensemble, subject-specific LDA classifiers were trained using leave-one-word-out cross-validation. For the generic ErrP classifier, EEGNet models were trained on data pooled across participants with leave-one-subject-out validation for 50 epochs using the Adam optimizer and an initial learning rate of 1e-4. Raw features were extracted from the 8-channel subset. We compared short (1.25 or 1.3 secs) and long (2 secs) feature window lengths for all classifiers, and two feature types (block-averaged and xDAWN spatial filtering enhanced with Riemannian geometry (RG) (41; 49)) for the LDA classifier.

The trained ErrP classifiers on error detection in the P300 speller were evaluated on data from the fake feedback phase of held-out participants in study C and the test phase of the rest of the studies in our dataset. The ensemble-based classification decision was based on a voting scheme, with the weight of each source user's classifier decision either assigned uniformly or based on the similarity between the source and novel users' data. The similarity-based weight of the source subject was computed using on the 4^{th} power of the inverse of the Kullback-Liebler divergence between source and novel users' EEG data and normalized across source users (50).

4.3.3 Statistical Analyses

Our statistical tests account for repeated measures from each participant to evaluate the within-subject changes in performance across conditions. Linear mixed effects (LME) models were estimated using maximum likelihood (*lmerTest* package in R) with Akaike information criterion (51) for comparisons. Character accuracy was the response variable, with participant as a random effect, fixes effects of classifier model, spatial filtering method, channels subset and downsampling decimation factor, and the interaction effects between classifier type and each of the other fixed effects. Estimated marginal means were determined after adjusting for the various factors and their interactions (*emmeans* package in R). Statistical significance was assessed using t-tests with Satterwaithe's method, $p < 0.05$. For the ErrP analysis, we used Wilcoxon signed-rank tests ($p < 0.05$) to assess the effect of EEG window length and classifier model on performance, with Bonferroni adjustments for pairwise comparisons.

5 Results

5.1 Data Exploration

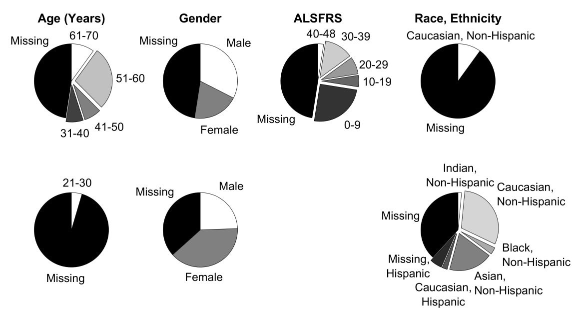


Figure 3: Summary of participant demographics of the current batch of our BCI dataset extracted from the .edf file metadata. ALSFRS-r, revised amyotrophic lateral sclerosis (ALS) Functional Rating Scale.

Figure 3 summarizes population-specific demographics, including the proportion of missing data. Most of the missing demographic data are from earlier studies where the information was not adequately preserved over time. All participants had to be at least 18 years to be in the study. While demographics trends cannot be fully assessed due to the missing data, anecdotal evidence indicates the age range of participants without ALS tends to be younger (18-30 years) and more racially/ethnically diverse when compared to participants with ALS. For studies with participants without disability, we primarily recruit from a university population for practical reasons to conduct online studies with fewer logistical hurdles prior to testing promising candidates in the ALS population; this also reflects participant trends in the literature (52; 53; 44).

We also analyzed characteristics of average P300 ERPs, which include peak amplitude and latency, Figure C3 (Appendix C.3). Consistent with the literature (54), we observed that P300 ERPs are generally low amplitude ($< 10\ \mu$V), exhibit variable latency between 200ms to 600ms with an expected mode around 300 ms, and are most prominent at fronto-central electrodes (Fz and Cz). The distribution of P300 ERP characteristics is similar across populations.

5.2 Target Character Estimation with P300 Classifiers

Character selection accuracy with participant-specific P300 classifiers are summarized by population in Figure 4 using boxplots to visualize the performance range across participants. Estimated marginal means interaction plots and summaries of statistical tests are provided in Appendix D.2.2. In general, statistically significant improvements in accuracy were observed with increased decimation factor and xDAWN filtering, while no significant difference were observed with between channel subsets, $p = 0.59$. The effect size of preprocessing was larger in participants without ALS compared to participants with ALS; this is likely because the former population had a broader range in participant performance levels allowing for a wider margin for improvement (see Figure C.2 for online performances). Overall, these results suggest that deep learning models for P300 classification

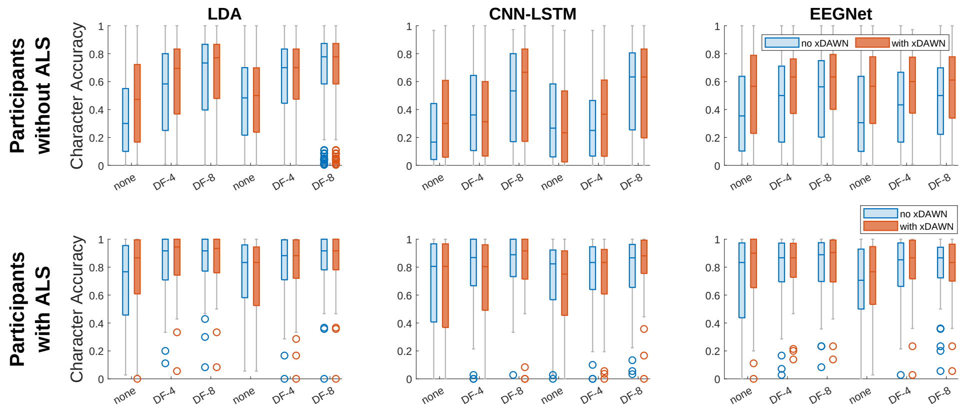


Figure 4: Box plots of P300 speller character accuracy with various signal preprocessing methods and classifier models. (a) Linear discriminant analysis (LDA); (b) convolutional neural network long short-term memory (CNN-LSTM) network (47); (c) EEGNet (46). Results are grouped by downsampling decimation factor (DF), channel subset, spatial filtering and participant population.

could potentially benefit with data preprocessing over minimally processed data. It should be noted that LDA demonstrates superior properties in totality in terms of model performance, training time and inference time when considering the time requirements for P300 ERP classification (Figure D1).

5.3 BCI Error Detection with ErrP Classifiers

Performance with generic EEGNet and subject-specific LDA ErrP classifiers of participants from the fake BCI feedback study are shown in Figure 5. Statistical results are summarized in Table **??**. For each model and input feature combination, features extracted from the long time windows achieved statistically significantly better performance relative to the short window, $p < 0.05$. For the long time window, the performance of the generic EEGNet classifiers was comparable to that of subject-specific LDA classifiers ($p < 0.51$). The rest of the ErrP analysis uses the best performing feature set for each classifier type, (LDA, xDAWN-RG, 2s) and (EEGNet, raw, 2s).

Figure 6 shows the precision and recall of the ensemble-based LDA and generic EEGNet ErrP classifiers. An appreciable number of classifiers predicted a single class. The LDA ensemble with uniforms weights almost entirely predicted a single class. While the LDA classifier ensemble with similarity-based weighting generated much fewer single class predictions, its performance was still

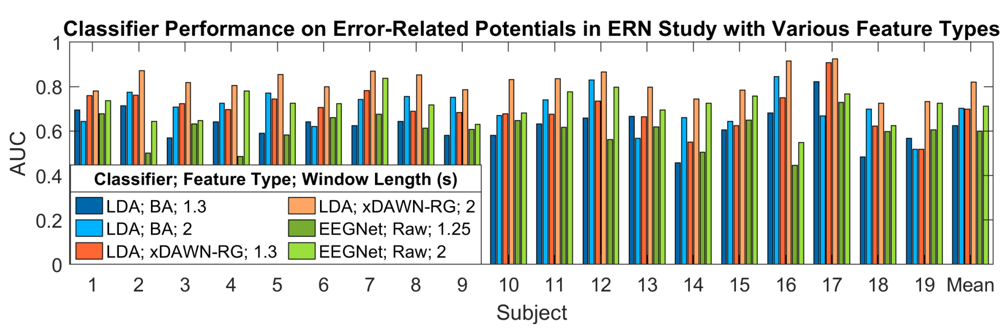


Figure 5: Areas under the receiver operating characteristics curve (AUC) of error-related potential classifiers for participants from the fake BCI feedback paradigm study (27). Subject-specific linear discriminant analysis (LDA) classifiers were trained with leave-one-word-out on block-averaged (BA, blue) or xDAWN-Riemannian geometry (RG) features (orange) extracted from 1.3s or 2.0s time windows. Generic EEGNet classifiers (green) were trained on raw features extracted from 1.25s or 2.0s windows with leave-one-subject out folds and applied to data from the held out participant.

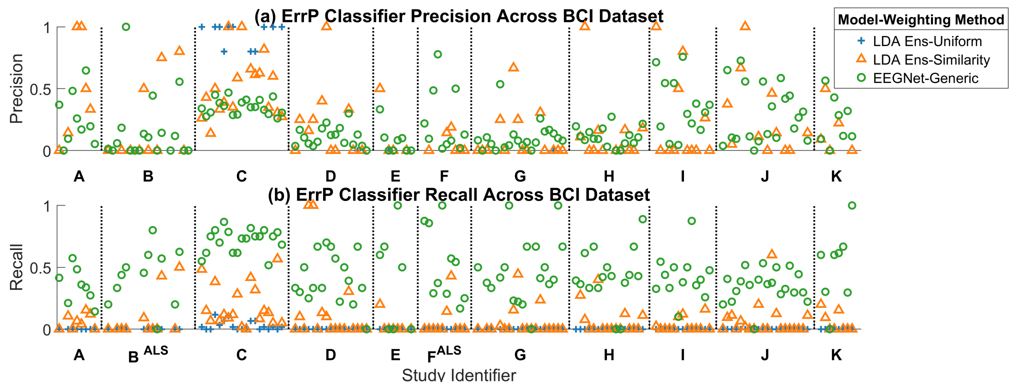


Figure 6: (a) Precision and (b) recall of error detection with error-related potential (ErrP) classifiers trained on data from the error-related negativity (ERN) study (Study C) and applied to target subjects segregated by study. Classifier models included an ensemble of subject-specific LDA classifiers using xDAWN-Riemannian geometry features and the generic EEGNet classifier using raw features on data pooled across source subjects, both extracted from 2-second window. Ensemble (Ens) voting used either with uniform or source-target similarity weighting of each classifier's decision in the ensemble.

poor. Similarly, EEGNet resulted in fewer single class predictions and generally had the highest recall, though its performance is not practical for robust error detection. Overall, the results further highlight challenges with generalizability of ErrP classifier to novel BCI users. Fine-tuning the generic EEGNet classifier to subject-specific data will be explored.

6 Discussion and Future Work

Our analyses presented here are not meant to be conclusive or comprehensive as the primary goals were to showcase the availability of our BCI dataset for analysis with open source tools, the broad user performance trends and reusability of our BCI dataset. In addition, these analyses enabled us to test data quality and refine our data engineering protocol to a near final version; thus, we limited the number of conditions for iterative dataset development and pilot testing to obtain feedback from different individuals independently analyzing the developing dataset. Our BCI dataset covers a wide range of experimental conditions that were investigated in the various P300 speller studies, potentially introducing additional confounds during data analysis. Nonetheless, diversity in the data conditions can serve as a test case of real-world data conditions inherent with variations. The large size and richness of our BCI dataset provides the opportunity for researchers to run a wide range of experiments where subsets of data can be carved out if similar conditions are needed.

We acknowledge the limitations of our work. which include missing demographic data and the younger skew of the participants without disability. We may also have missed relevant BCI datasets in our literature review. Our BCI data spans over a decade of BCI research and this extended time period creates challenges with data curation beyond missing data; our dataset development had to account for different BCI2000 versions over the years, different BCI2000 builds across institutions and varied data archiving formats across experimenters over the years. In recent years, we have mitigated the issue of missing demographic data by standardizing our data collection instruments and migrating from paper to electronic forms for participant consent and demographic data collection.

It is possible that our BCI data dictionary may need to be updated once the IEEE P2731 standards for BCI data storage and sharing are finalized. Since we have defined our BCI data dictionary based on the proposed BCI data levels and optimized our data reformatting protocol, we can easily adapt our BCI dataset to be compliant with any changes in the official IEEE P2731 BCI data standards.

Our large, diverse, machine learning ready P300 speller dataset of 170 participants provides an open resource to facilitate robust and well-powered offline analyses in BCI research. To the best of our knowledge, our final BCI dataset would represent the largest collection of publicly available P300 speller data from individuals with ALS in a single dataset (39 participants with ALS in the current BCI dataset). Dataset development from the rest of the collection of our source BCI data files (from

163 participants, including 10 participants with ALS) is still going and will be ready in the next couple of months after undergoing data quality control. We plan to publicly release the currently available BCI dataset under a creative commons license via the open PhysioNet repository.

Acknowledgments and Disclosure of Funding

The dataset development was funded by a grant from the National Institutes of Health (R21DC018347-02S1). The authors would like to thank Mr. Thomas Napoles, Ms. Katie Kilroy and Dr. John Board at Duke University for their technical support during the preparation of this dataset.

References

[1] J. R. Wolpaw and E. W. Wolpaw, "Brain-computer interfaces: Something new under the sun," *Brain-Computer Interfaces: Principles and Practice*, vol. 14, 2012.

[2] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, 2018.

[3] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *Journal of Neural Engineering*, vol. 16, no. 3, p. 031001, 2019.

[4] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. Mcalpine, and Y. Zhang, "A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers," *Journal of Neural Engineering*, vol. 18, no. 3, p. 031002, 2021.

[5] J. Lee, K. Won, M. Kwon, S. C. Jun, and M. Ahn, "CNN with large data achieves true zero-training in online P300 brain-computer interface," *IEEE Access*, vol. 8, pp. 74 385–74 400, 2020.

[6] L. Bianchi, A. Antonietti, G. Bajwa, R. Ferrante, M. Mahmud, and P. Balachandran, "A functional BCI model by the IEEE P2731 working group: data storage and sharing," *Brain-Computer Interfaces*, vol. 8, no. 3, pp. 108–116, 2021.

[7] W. Speier, C. Arnold, and N. Pouratian, "Integrating language models into classifiers for BCI communication: A review," *Journal of Neural Engineering*, vol. 13, no. 3, p. 031002, 2016.

[8] J. S. Brumberg, K. M. Pitt, A. Mantie-Kozlowski, and J. D. Burnison, "Brain–computer interfaces for augmentative and alternative communication: A tutorial," *American Journal of Speech-Language Pathology*, vol. 27, no. 1, pp. 1–12, 2018.

[9] G. Schalk and J. Mellinger, *A practical guide to brain–computer interfacing with BCI2000: General-purpose software for brain-computer interface research, data acquisition, stimulus presentation, and brain monitoring*. Springer Science & Business Media, 2010.

[10] P. Stegman, C. S. Crawford, M. Andujar, A. Nijholt, and J. E. Gilbert, "Brain-computer interface software: A review and discussion," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 2, pp. 101–115, 2020.

[11] C. Easttom, L. Bianchi, D. Valeriani, C. S. Nam, A. Hossaini, D. Zapała, A. Roman-Gonzalez, A. K. Singh, A. Antonietti, and G. Sahonero-Alvarez, "A functional model for unifying brain computer interface terminology," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 2, pp. 91–96, 2021.

[12] B. Blankertz, K. R. Muller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schlogl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schroder, and N. Birbaumer, "The BCI Competition 2003: Progress and perspectives in detection and discrimination of EEG single trials," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1044–51, 2004.

[13] B. Blankertz, K. R. Muller, D. J. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlogl, G. Pfurtscheller, R. Millan Jdel, M. Schroder, and N. Birbaumer, "The BCI Competition III: Validating alternative approaches to actual BCI problems," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 153–9, 2006.

[14] C. Guger, S. Daban, E. Sellers, C. Holzner, G. Krausz, R. Carabalona, F. Gramatica, and G. Edlinger, “How many people are able to control a P300-based brain-computer interface (BCI)?” *Neuroscience Letters*, vol. 462, no. 1, pp. 94–8, 2009.

[15] L. Citi, R. Poli, and C. Cinel, “Documenting, modelling and exploiting P300 amplitude changes due to variable target delays in donchin’s speller,” *Journal of Neural Engineering*, vol. 7, no. 5, p. 056006, 2010.

[16] M. S. Treder, N. M. Schmidt, and B. Blankertz, “Gaze-independent brain-computer interfaces based on covert attention and feature attention,” *Journal of Neural Engineering*, vol. 8, no. 6, p. 066003, 2011.

[17] F. Aloise, P. Aricò, F. Schettini, A. Riccio, S. Salinari, D. Mattia, F. Babiloni, and F. Cincotti, “A covert attention P300-based brain-computer interface: Geospell,” *Ergonomics*, vol. 55, no. 5, pp. 538–51, 2012.

[18] L. Acqualagna and B. Blankertz, “Gaze-independent BCI-spelling using rapid serial visual presentation (RSVP),” *Clinical Neurophysiology*, vol. 124, no. 5, pp. 901–908, 2013.

[19] A. Riccio, L. Simione, F. Schettini, A. Pizzimenti, M. Inghilleri, M. O. Belardinelli, D. Mattia, and F. Cincotti, “Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis,” *Frontiers in Human Neuroscience*, vol. 7, p. 732, 2013.

[20] K. Won, M. Kwon, M. Ahn, and S. C. Jun, “EEG Dataset for RSVP and P300 Speller Brain-Computer Interfaces,” *Scientific Data*, vol. 9, no. 1, p. 388, 2022.

[21] J. M. Cedarbaum *et al.*, “The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function,” *Journal of the Neurological Sciences*, vol. 169, no. 1-2, pp. 13–21, 1999.

[22] A. Kübler, E. M. Holz, A. Riccio, C. Zickler, T. Kaufmann, S. C. Kleih, P. Staiger-Sälzer, L. Desideri, E.-J. Hoogerwerf, and D. Mattia, “The user-centered design as novel perspective for evaluating the usability of BCI-controlled applications,” *PloS One*, vol. 9, no. 12, p. e112392, 2014.

[23] A. Kübler, F. Nijboer, and S. Kleih, “Hearing the needs of clinical users,” *Handbook of clinical neurology*, vol. 168, pp. 353–368, 2020.

[24] “BCI2000 Contributions: EyetrackerLogger.” [Online]. Available: https://www.bci2000.org/mediawiki/index.php/Contributions:EyetrackerLogger

[25] C. S. Throckmorton, D. B. Ryan, B. Hamner, K. Caves, K. A. Colwell, E. W. Sellers, and L. M. Collins, “Towards clinically acceptable bci spellers: Preliminary results for different stimulus selection patterns and pattern recognition techniques.” 2010.

[26] N. A. Gates, C. K. Hauser, and E. W. Sellers, “A longitudinal study of P300 brain-computer interface and progression of amyotrophic lateral sclerosis,” in *International Conference on Foundations of Augmented Cognition*. Springer, 2011, Conference Proceedings, pp. 475–483.

[27] B. O. Mainsah, K. D. Morton, L. M. Collins, E. W. Sellers, and C. S. Throckmorton, “Moving away from error-related potentials to achieve spelling correction in P300 spellers,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 5, pp. 737–743, 2015.

[28] B. O. Mainsah, K. A. Colwell, L. M. Collins, and C. S. Throckmorton, “Utilizing a language model to improve online dynamic data collection in P300 spellers,” *Ieee Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 4, pp. 837–846, 2014.

[29] B. O. Mainsah, L. M. Collins, K. A. Colwell, E. W. Sellers, D. B. Ryan, K. Caves, and C. S. Throckmorton, “Increasing BCI communication rates with dynamic stopping towards more practical use: an ALS study,” *Journal of Neural Engineering*, vol. 12, no. 1, 2015.

[30] B. Mainsah, K. Morton, L. Collins, and C. Throckmorton, “Extending language modeling to improve dynamic data collection in ERP-based spellers.” in *6th International Brain-Computer Interface Conference*, 2014, Conference Proceedings.

[31] D. Kalika, L. Collins, K. Caves, and C. Throckmorton, "Fusion of P300 and eye-tracker data for spelling using BCI2000," *Journal of Neural Engineering*, vol. 14, no. 5, p. 056010, 2017.

[32] B. Mainsah, G. Reeves, L. Collins, and C. Throckmorton, "Optimizing the stimulus presentation paradigm design for the P300-based brain-computer interface using performance prediction," *Journal of Neural Engineering*, vol. 14, no. 4, p. 046025, 2017.

[33] B. Mainsah, D. Kalika, L. Collins, S. Liu, and C. Throckmorton, "Information-based adaptive stimulus selection to optimize communication efficiency in brain-computer interfaces," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[34] D. B. Ryan, K. A. Colwell, C. S. Throckmorton, L. M. Collins, K. Caves, and E. W. Sellers, "Evaluating brain-computer interface performance in an ALS population: checkerboard and color paradigms," *Clinical Electroencephalography and Neuroscience*, vol. 49, no. 2, pp. 114–121, 2018.

[35] G. Pires, M. Castelo-Branco, C. Guger, and G. Cisotto, "Error-related potentials: Challenges and applications," *Frontiers in Human Neuroscience*, vol. 16, 2022.

[36] B. Kemp and J. Olivan, "European data format 'plus'(EDF+), an EDF alike standard format for the exchange of physiological data," *Clinical Neurophysiology*, vol. 114, no. 9, pp. 1755–1761, 2003.

[37] "European Data Format." [Online]. Available: https://www.edfplus.info/

[38] "Racial and Ethnic Categories and Definitions for NIH Diversity Programs and for Other Reporting Purposes, Notice Number: NOT-OD-15-089," 2015. [Online]. Available: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-089.html

[39] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, and P. E. Bourne, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[40] D. J. Krusienski, E. W. Sellers, F. Cabestaing, S. Bayoudh, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "A comparison of classification techniques for the P300 speller," *Journal of Neural Engineering*, vol. 3, no. 4, pp. 299–305, 2006.

[41] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN algorithm to enhance evoked potentials: Application to brain-computer interface," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 8, pp. 2035–43, 2009.

[42] "pyRiemann: Biosignals classification with Riemannian geometry." [Online]. Available: https://pyriemann.readthedocs.io/en/latest/index.html

[43] L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–23, 1988.

[44] J. Kalra, P. Mittal, N. Mittal, A. Arora, U. Tewari, A. Chharia, R. Upadhyay, V. Kumar, and L. Longo, "How visual stimuli evoked p300 is transforming the brain-computer interface landscape: A prisma compliant systematic review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[45] "User Reference: P300 Classifier Methods." [Online]. Available: https://www.bci2000.org/mediawiki/index.php/UserReference:P300ClassifierMethods

[46] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, 2018.

[47] O. Tal and D. Friedman, "Using recurrent neural networks for P300-based brain-computer interface." in *Proceedings of the 7th Graz Brain-Computer Interface Conference*, 2017.

[48] " TheArmy Research Laboratory (ARL) EEGModels project: A Collection of Convolutional Neural Network (CNN) models for EEG signal processing and classification." [Online]. Available: https://github.com/vlawhern/arl-eegmodels

[49] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain-computer interface classification by Riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–8, 2012.

[50] A. M. Azab, L. Mihaylova, K. K. Ang, and M. Arvaneh, "Weighted transfer learning for improving motor imagery-based brain–computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1352–1359, 2019.

[51] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer Science & Business Media, 2003.

[52] B. S. Eddy, S. C. Garrett, S. Rajen, B. Peters, J. Wiedrick, D. McLaughlin, A. O'Connor, A. Renda, J. E. Huggins, and M. Fried-Oken, "Trends in research participant categories and descriptions in abstracts from the international bci meeting series, 1999 to 2016," *Brain-Computer Interfaces*, vol. 6, no. 1-2, pp. 13–24, 2019.

[53] B. Z. Allison, A. Kübler, and J. Jin, "30+ years of p300 brain–computer interfaces," *Psychophysiology*, vol. 57, no. 7, p. e13569, 2020.

[54] R. van Dinteren, M. Arns, M. L. Jongsma, and R. P. Kessels, "P300 development across the lifespan: a systematic review and meta-analysis," *PloS one*, vol. 9, no. 2, p. e87347, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sections 3.3, 4 and 5.
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See 3.1.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] See Sections 3.1, 3.2.1,5.1.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Instructions are included in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Sections 4.3.3 and 5
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We are the creators of the BCI dataset. We cite the use of existing code and models, see Section 4.
   (b) Did you mention the license of the assets? [Yes] The plan is to release the final dataset under a creative commons license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Sections 3.1 and 3.2.1

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 3.1

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] The dataset includes data from several studies. A general description of the BCI protocol is provided in Section 3.2.3 and detailed in related publications.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] See section 3.1.

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Section 3.2.1

Appendix

Table of Contents

A Datasheet for Dataset

A.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Brain–computer interfaces (BCIs) have wide-ranging applications that are solutions for replacing neural output that has been lost because of injury and disease, such as individuals with late-stage amyotrophic lateral sclerosis (ALS). Current BCIs have relatively low communication rates due to the inherent limitations associated with processing inherently noisy data and highly variable neural signal components of interest to extract the relevant information that is needed to control the BCI. Thus, improving BCI communication efficiency is an area of significant research interest. Part of the development process for any BCI algorithm involves performing simulations with electroencephalography (EEG) data collected from previous BCI studies to pre-assess various BCI algorithms or strategies under consideration and selecting promising candidates for real-world testing. Acquiring BCI data is time-consuming and expensive; thus, most BCI research groups rely on publicly available datasets to obtain the necessary data to perform simulations with EEG data, rather than conducting real-time BCI studies in-house. However, current publicly available BCI datasets have a limited number of participants, have an under-representation of target BCI end users, and mostly use a proprietary file format. Also, there is a lack of serial data collected over several hours and days of BCI use that are needed for simulating long-term evaluation of BCI algorithms. Based on research support from the National Institutes of Health for 10-plus years, we have acquired a large amount of single- and multi-session data from P300-based BCI speller studies with abled-bodied individuals and individuals with ALS under a wide range of experiment conditions. Guided by FAIR principles, we will perform data curation, data cleaning, and data engineering to transform proprietary data files into an open and nonproprietary file format; package the transformed files into a machine-readable dataset with metadata and documentation; and make this transformed dataset publicly available via an open-source repository.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? This BCI dataset was created by Boyla O. Mainsah (Principal Investigator), Leslie M. Collins, Chance Fleeting (Technical lead), Thomas Balmat (Dataset developer), Wei Wu, Eric J. Qi and William Luqiu at Duke University, and Eric Sellers at East Tennessee State University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The BCI studies were funded by the National Institutes of Health under a grant supplement administered by the National Institute on Deafness and Other Communication Disorders (R21DC018347-02S1).

A.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. The dataset consists of files in European Data Format plus (.edf+) format containing data from adult human participants generated during P300-based BCI experiments. Each participant completed one or multiple BCI experiment sessions, during which they received brief instruction and performed copy-spelling tasks with the P300 speller.

How many instances are there in total (of each type, if appropriate)? The current dataset contains data from 170 participants, including 39 participants with ALS who completed one or multiple BCI experiment sessions.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable). The dataset is a sample of EEG data from adult human subjects sampled from the larger populations of (1) people without neuromuscular disability and (2) people with amyotrophic lateral sclerosis. It is not necessarily representative of the larger set, due to the geographic limitations of recruiting human subjects for in-person experiments.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The data dictionary for the .edf+ file is detailed in Table A1

Table A1: Our BCI data dictionary for the file header and data records in .edf files based on developing IEEE P2731 standards for brain-computer interface (BCI) data storage and sharing.

(a) EDF file header.

Identification Field	Sub-field	Modified Sub-field Label
Patient	Patient code	Subject number
	Sex	
	Date of birth	01-JAN-YYYY Age (years) = 2020 minus YYYY
	Patient name	<*Race*>_<*Ethnicity*>_<*ALS Status*>
Recording	Start date	01-JAN-2020
	Hospital admin code	Study identifier
	Technician	Session number
	Equipment code	Equipment model

(b) EDF data

BCI Data Level		Data Record	Data Label(s)
Level 0: Biosignals		EEG signals	EEG_<*channelName*>
	Eye tracker signals	Eye data validity	ET<Left/Right>EyeValid
		Eye gaze position	ET<Left/Right>EyeGaze<X/Y>
		Eye position	ET<Left/Right>EyePos<X/Y>
		Eye distance	ET<Left/Right>EyeDist
		Pupil size	ET<Left/Right>PupilSize
Level 1: BCI Training		Character trial events	PhaseInSequence
		Stimulus events	StimulusBegin, StimulusType
		Character presentation events	<*Character*>_<*row#*>_<*column#*>
		Target character	CurrentTarget
Level 2: BCI Feedback		Predicted target character	Selected<Target/Row/Column>
		Presented BCI feedback	DisplayResult, FakeFeedback

For sub-field and data labels, <*option*> in italics indicates a variable substring within the angle brackets and <option1/…/optionN> in solid indicates a variable substring from the set {option1,..., optionN}. E.g., ET<Left/Right>EyeDist indicates two options, ETLeftEyeDist and ETRightEyeDist, for the eye distance label fro the eye tracker (ET). Most of the EDF+ data labels we used are derived from parameter definitions in BCI2000 *.dat* files (9).

Is there a label or target associated with each instance? If so, please provide a description.

1. Stimulus type labels, time-locked to stimulus presentation onset, can be used to train a P300 event related potential classifier.

2. True and predicted target character labels, time-locked to onset of displayed spelling trial feedback, can be used to train an error-related potential classifier.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text. Demographic data are missing if they were not requested from the participant or unavailable due to lack of preservation. Eye tracker data are missing if they were either not collected during an experiment (complete absence) or are invalid due to inability to detect the eye during an experiment (intermittent).

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. In each study, each participant is assigned an identification number, which is consistent.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. The data for each participant are split according to the training and test blocks of the BCI experiment session(s). During the training (or calibration) block, copy-spelling of predefined words is performed with no BCI feedback to collect labeled EEG data to train a BCI classifier. During the test block, the trained BCI classifier is applied to perform copy-spelling of predefined words with BCI feedback to evaluate a BCI algorithm or strategy.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. EEG recordings may contain power line noise and movement artifact (e.g., eye blink, head movement, etc.). EEG signal preprocessing includes bandpass filtering, and if applicable, notch filtering (to remove 60 Hz line noise). Parameters of signal filters, if applied, are specified in the *prefiltering* field of the .edf+ file. Eye tracker may be unable to acquire data when the eye is not detected (e.g., during eye blink); a data record to indicate validity of eye gaze data is included.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? This dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description. No, all personal identifiable information and protected health information have been removed from the dataset.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. Participant data may include self-reported demographic information (age, race, ethnicity and sex), as well as ALS diagnosis and ALSFRS-r score abstracted from medical records.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how. No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?If so, please provide a description. Yes, participant data may include self-reported demographic information (race, ethnicity, sex, age) and health data abstracted from medical records (ALS diagnosis, ALSFRS-r score). IRB approval was obtained to release limited data abstracted from the medical records.

A.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how. The EEG data were directly observable. Time-stamps of BCI actions were provided by the BCI2000 platform (9). Self-reportated demographic information is sometimes available.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated? EEG data were recorded using gel-based passive electrodes or dry active electrodes connected to either one 16-channel or two 16-channel gUSBamp (g.tec medical engineering GmbH) biosignal amplifiers with a 10-20 electrode montage. An electrode impedance check was conducted to ensure low impedance prior to EEG signal recording. Eye tracker data were recorded using a Tobii Pro X2-30 (Tobii AB) infrared eye tracker synchronised to EEG data collection during BCI use. The eye tracker was calibrated for each participant prior to BCI use.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Participants were recruited based on their interest following study advertisements on campus or via an affiliated clinic.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? IRB approved personnel on the BCI project were involved in the data collection process, which included graduate students, research staff and faculty at the university.

Over what timeframe was the data collected? Does this timeframe match the creation time-frame of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time-frame in which the data associated with the instances was created. Per institutional regulatory requirements, we are not allowed to disclose the time frame when data were collected. Publications related to the BCI studies will be referenced.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. BCI studies were approved by the Institutional Review Boards. The dataset release is presently undergoing institutional regulatory review to ensure personal identifiable data have been removed and the data transfer agreement with PhysioNet is finalized.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? The data were collected directly from the participants or (when applicable) abstracted from medical records.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself. Participants were recruited via flyers, electronic bulletins and letters, which included a description about the BCI study. The consent form included a description of the BCI experiment and the data collection process. Participants also received a brief description of the BCI experiment prior to performing the BCI spelling tasks.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. All participants provided informed consent prior to data collection. Participant were made aware that their data collected during the experiment may be made available to the public, with identifiable information removed to preserve privacy.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). Yes, the consent form contained language that indicated that participants could withdraw consent at any point during their experiment and be compensated for their time up to withdrawal of consent. Participants recruited via health systems were informed that withdrawal of care would have no impact on their patient care. The option to withdraw consent was also repeated prior to beginning an experiment session. Furthermore, participants were informed that they could contact the Chair of the Institutional Review Board for the Protection of Human Subjects for any question concerning their rights as a research participant, and that they could contact the investigators with any questions. Contact details in the form of phone numbers were provided.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. No

A.4 Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section. EEG signal preprocessing includes bandpass filtering, and if applicable, notch filtering (to remove 60 Hz line noise). Parameters of signal filters, if applied, are specified in the prefiltering field of the .edf file.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. No, raw EEG signals are not saved as EEG signals are bandpass and notch filtered at the biosignal amplifier stage prior to storage in the source (.dat) files in BCI2000.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point. EEG signals are filtered at the biosignal amplifier stage with proprietary code prior to storage.

A.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description. This dataset is a curation of data from several P300 speller studies that have been peer-reviewed and published, as well as related unpublished studies.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. A list of the related publications is provided in Table 2 in the main text.

What (other) tasks could the dataset be used for? The dataset can be used for various machine learning and signal processing tasks related to EEG data from P300 speller use, such as developing a BCI classifier, novel EEG data representations and signal filtering techniques.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms? Dataset consumers should note the available demographic information, which includes information related to ALS for some participants, and be careful about making generalizations, particularly to out-of-sample populations.

Are there tasks for which the dataset should not be used? If so, please provide a description. The dataset should not be used for person identification from EEG data or to otherwise infringe on the rights of the participants included in the dataset.

A.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. Yes, the BCI dataset will be made available on PhysioNet, a publicly available repository.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? When released, the dataset will be downloaded from the website physionet.org.

Does the dataset have a digital object identifier (DOI)? All datasets hosted on PhysioNet are assigned a DOI.

When will the dataset be distributed? The dataset will be distributed once submitted and the PhysioNet editorial review is completed, which is expected to occur in the next couple of months.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. The dataset will be distributed under a Creative Commons license (https://creativecommons.org/licenses/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. No.

A.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The dataset will be hosted on PhysioNet.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Please contact boyla.mainsah@duke.edu for matters concerning this dataset.

Is there an erratum? If so, please provide a link or other access point. No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)? The dataset may be updated periodically to correct labeling errors or add new instances at an undetermined schedule. The current dataset version with dates and changes will be noted.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced. No, fixed limits on the retention of the data are not imposed.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers. Older versions of the dataset will not be supported. The current dataset version with dates and changes will be noted once available via PhysioNet.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description. No external contributions to the dataset will be supported.

B Author Statement

The authors bear all responsibility in case of violation of rights. The dataset will be issued under a Creative Commons license.

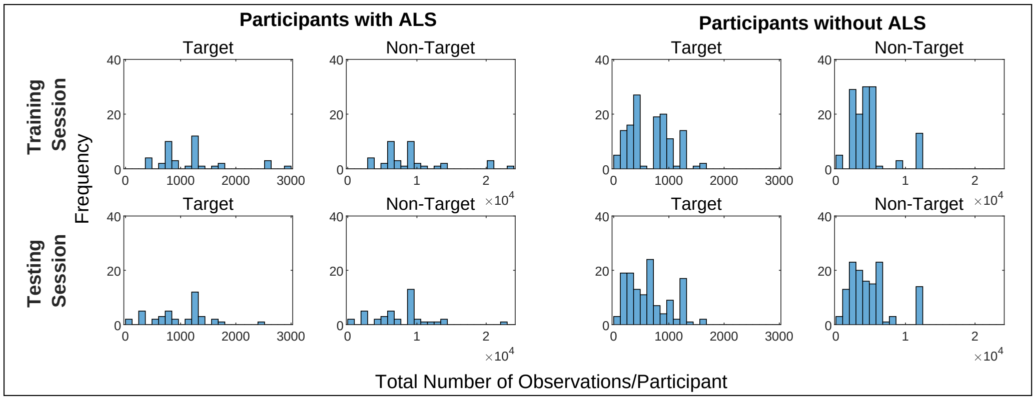


Figure C1: Target/non-target

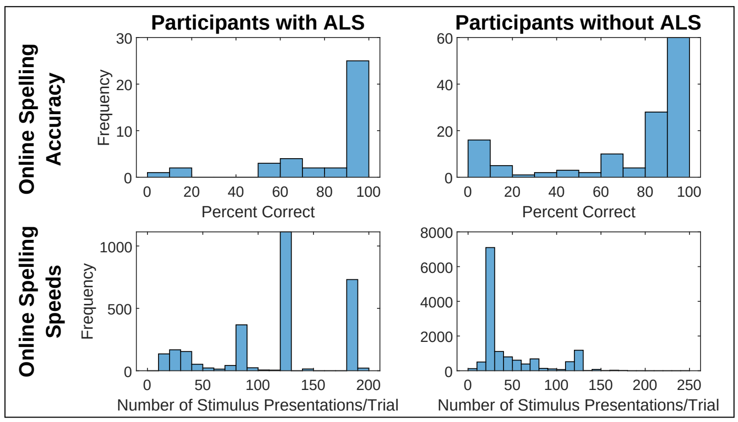


Figure C2: Online performance

C Data Exploration

C.1 Data type distribution

C.2 Online Performance

C.3 Evoked Potentials

D Data Analysis

D.1 Signal Preprocessing

Signal preprocessing methods, downsampling factors, and electrode montages were compared. XDAWN preprocessing was performed using the pyriemann package in Python, with four filters per class. Downsampling was performed in a channel-wise manner via the SciPy package in Python. The 8-channel electrode montage was {Fz, Cz, P3, Pz, P4, PO7, PO8, Oz} and the 16-channel montage was {F3, Fz, F4, T7, C3, Cz, C4, T8, CP3, CP4, P3, Pz, P4, PO7, PO8, Oz}.

D.2 P300 Analysis

D.2.1 Deep Learning Model Architecture

EEGNet and CNN-LSTM models were trained in Tensorflow. The EEGNet-8,2 architecture (46) with a dropout rate of 0.5 was used. The CNN-LSTM architecture was a small version with 30 hidden

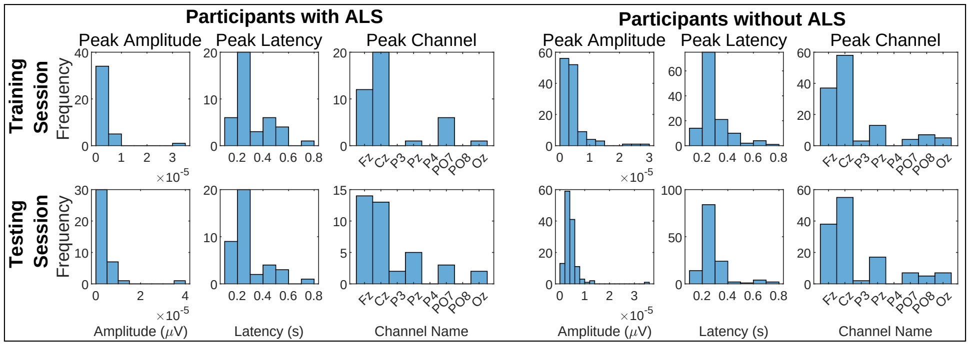


Figure C3: P300 ERP.

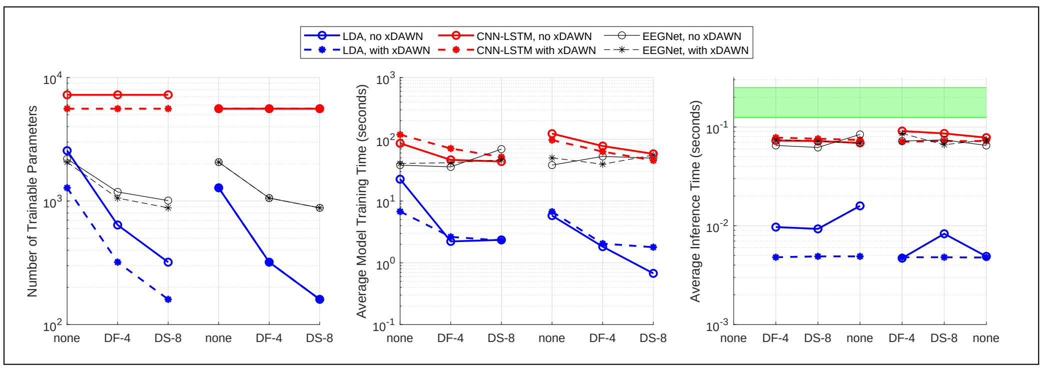


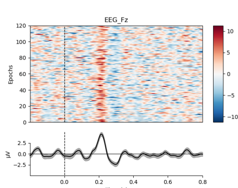
Figure D1: Model Comparison. Green region is the range of typical inter-stimulus interval times.

layers (47). Models used a tanh activation function in the last layer. Table **??** shows the mean training and testing times for these models.

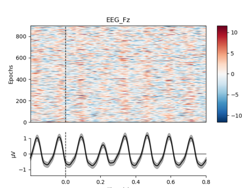
D.2.2 Statistical Analyses

P300 speller target characters were predicted by summing up classifier scores associated with each character on the board and finding the character that maximized this sum for the current spelling trial. The character accuracy was calculated based on these predictions. From Figure D3 and Table D1, it is evident that downsampling had a strong effect on performance, with a downsampling factor of 8 consistently leading to improved accuracy. Similarly, xDAWN preprocessing consistently improved performance, particularly for EEGNet (Figure D3c). The CNN-LSTM model had the lowest performance overall, with LDA and EEGNet being comparable in performance. Finally, the choice of channel set did not seem to impact performance for any model.

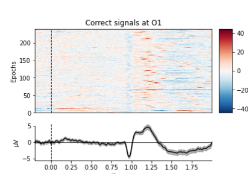
The LDA models with 16 channels, no downsampling and no xDAWN filtering was the baseline. With LDA as reference, CNN-LSTM performance was statistically significantly worse ($p < 0.01$), while performance with EEGNet was comparable ($p = 0.07$). The effect conferred by a combination



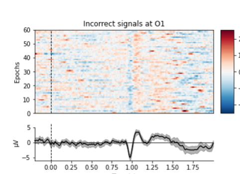
(a) Target stimuli



(b) Nontarget stimuli



(c) Correct feedback



(d) Incorrect feedbackxd

Figure D2: Three simple graphs

Table D1: Contributors to P300 Classification Accuracy by Factor.

Factor	Contrast	p-val
Model	CNN-LSTM vs. EEGNet	$< \mathbf{0.01}$
xDAWN	Without vs. With	$< \mathbf{0.01}$
Channel subset	8 vs. 16	0.49
Decimation factor	none vs. 4	$< \mathbf{0.01}$
	none vs. 8	$< \mathbf{0.01}$

Statistically significant differences in performance are bolded.

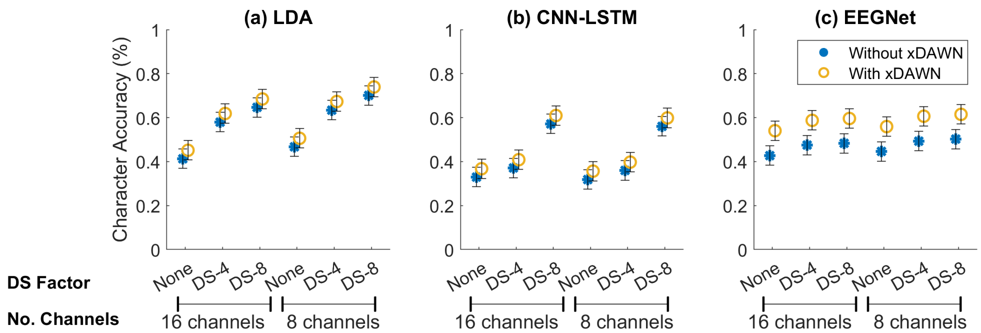


Figure D3: Estimated marginal means for character selection accuracy across all levels of fixed factors. Fixed effect terms were: classifier model (Linear discriminant analysis (LDA), CNN-LSTM, EEGNet), spatial filtering (no xDAWN and xDAWN), the number of input EEG channels (8 and 16), EEG downsampling decimation factor (none, 4, and 8). Interaction effects were included between model and each of the other fixed effects. Error bars denote lower and upper 95% confidence levels.

of factors was model dependent: all interaction effects were statistically significant ($p < 0.01$), except for CNN-LSTM with downsampling factor of 8 and xDAWN processing. Overall, these results suggest that deep learning models for P300 classification could potentially benefit with data preprocessing over the minimally processed data.

D.3 Error-Related Potential Analysis

The statistical analyses in Table **??** relate to the results in Figure 5. Overall, among the explored parameters, using the longer EEG window resulted in improved performance regardless of feature or model choice. With a 2s window, the xDAWN-RG features with an LDA classifier performed statistically significantly better than the other feature-model combinations.