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

Anonymous Author(s)
Affiliation
Address
email

Abstract

The P300-based brain-computer interface (BCI) is one of the most commonly researched BCIs for communication. We present an open large BCI dataset of 159+ participants that provides data in an enriched and standardized format, with study metadata, demographics and BCI data levels that align with developing IEEE BCI data standards to facilitate reusability in BCI research. Our BCI dataset, curated from our previous P300 speller studies, encompasses a wide range of user performance levels and experimental conditions, including longitudinal data for simulated extended BCI use. To address the current under-representation of target end users in BCI data, our dataset includes data from 29+ individuals with amyotrophic lateral sclerosis (ALS), a target BCI end user population. We demonstrate the broad utility of our dataset with experiments of machine learning tasks for character selection and error detection in P300 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 application, 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 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 or strategies under consideration prior to, or even in place of, online testing, to identify a few promising candidates for subsequent online testing. In particular, the performance of algorithms for BCIs intended for long-term use may be best captured via simulations with longitudinal BCI data. However, a core challenge to ongoing BCI research is the relatively low-resource environment, especially when compared to datasets for machine learning applications such as computer vision and natural language processing. BCI data acquisition is time-consuming and expensive, particularly if it is necessary to collect data from the same users over several hours or several days of BCI use. Most BCI research groups often obtain data to perform BCI simulation from publicly available datasets, rather than collecting data in-house (e.g., (2)).

Need for Big Data in BCI Research. Due to high variability across individuals, the standard approach for BCI machine learning is user-specific. Alternatively, data from other BCI users can be leveraged for training generalized models, pretraining models to facilitate transfer learning or

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

analyzed in aggregate to identify performance trends or users with similar profiles. iven the successes in other fields, recent years have seen an increased interest in using deep learning for BCI applications (3; 4; 5). Such models benefit from learning from a large and diverse user dataset (6), 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 (**??**, 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 studies with target end users to validate BCI algorithms (7; 8) and these data are rarely disseminated. It is vital to perform simulations with EEG 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 data definitions, lack of or limited documentation in some cases, and differences in file format and data content across various BCI datasets (9). Most BCI research groups use the open-source BCI2000 software platform(10; 11), 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 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 pdf file, creating issues with portability (9). The IEEE P2731 working group (WG) is currently developing new standards for a unified BCI terminology (12), including for BCI data storage and sharing (9).

Our work focuses on the P300 speller, a widely researched BCI for communication. We have accumulated a large collection of BCI data from 10+ years of BCI research that includes studies with participants with ALS. To address the current gaps in current BCI research data, we have curated and developed a highly structured BCI dataset, with our our key contributions in this work are summarized as follows:

1. We introduce a new BCI data dictionary for the open file format, European Data Format plus (EDF+), that leverages its file specifications to embed study metadata, participant demographics and BCI data levels within files to align with currently developing IEEE standards for BCI data storage and sharing.
2. We present a large, diverse BCI dataset of 159+ participants, including 29 participants with ALS. Our BCI dataset encompassing a wide range of user performance levels and experimental conditions. When dataset development is finalized, our BCI dataset will contain 300+ participants, including 49 participants with ALS. The final BCI dataset will be made publicly available and will be unique in its size, coherency, diversity, and inclusion of target BCI end users.
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 where 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 EEG data for relevance to our BCI dataset. Table 1 summarizes the characteristics of BCI datasets we identified from the literature search (13; 14; 15; 16; 17; 18; 19; 20; 21). Only one study included participants with ALS (20). Most datasets have between 1-13 participants, except (21) with 55 participants. All but one of the datasets provide data files in binary MATLAB format. (16) provides European Data Format files, the earlier version of the file format used in our BCI dataset. Some BCI datasets do not provide user-specific demographics but summary statistics are available in a related publication

Dataset	N	ALS Population	No. of Sessions	Stimulus Paradigm(s)	File Format	User-specific Demographics
Blankertz et al., 2004 (13)	1	No	3	RC	MAT	-
Blankertz et al., 2006 (14)	2	No	5	RC	MAT	-
Guger et al., 2009(15)	10	No	1	RC, SC	MAT	-
Citi et al., 2010 (16)	12	No	1	RC	EDF	-
Treder et al., 2011 (17)	13	No	1	Regional	MAT	-
Aloise et al., 2012 (18)	10	No	1	Center, RC	MAT	{Age, Sex}, In a separate file
Acqualagna et al., 2013 (19)	12	No	1	RSVP	MAT	-
Won et al., 2022 (21)	55	No	1	RC	MAT	-
Riccio et al., 2013 (20)	8	Yes	1	RC	MAT	{Age, Sex, ALSFRS-r, Onset}, Embedded
Our dataset[†]	159+	Yes	1, 2-5	RC, CB, RD, PB, AD	EDF+	{Age, Sex, Race/Ethnicity, ALSFRS-r}, Embedded

Abbreviations: AD, Adaptive; BI, Brain Invaders; CB, Checkerboard; EDF, European Data Format; MAT, binary MATLAB files; MS, Multi-session; PB, Performance-Based; RD, Random; RC, Row-Column; RSVP, rapid serial visual presentation; SET, MATLAB EEGLAB .set files; SS, Single Session.

[†]Dataset in batches.

Table 1: Comparison of Publicly Available Visual P300 BCI Speller Datasets and Our Dataset.

(15; 19; 17; 21). One study includes user-specific demographics in a separate file (18). Similar to our BCI dataset, (20) embeds demographics in the same file as the BCI data.

Compared to similar public BCI datasets, our proposed P300 speller dataset is unique in having the following attributes: a) open files that provide data in an enriched and standardized format to facilitate reusability; b) a sufficiently large number of participants (159+) 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, 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. 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.

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 (22) with a score range of 0 to 48, where a lower value indicates a higher degree of functional impairment.

3.2.2 Technical Setup

Data were recorded using BCI2000 (10). For participants without ALS, a 9×8 (number of rows $\times$ number of columns) 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. EEG signals were collected non-invasively at 256 Hz using gel-based passive electrodes or dry active 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 prior to EEG signal recording (generally < 40 kΩ). Raw EEG data were bandpass filtered (0.5 - 30 Hz) and where applicable, notch filtered (60 Hz) to remove electrical noise; this was done internally in the biosignal amplifier so all signals are frequency filtered. For hybrid BCIs, 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 of the Tobii eye tracker data filter in BCI2000 (23).

3.2.3 Experiment Protocol

During an experiment session, 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. A user is presented with a set of choices on a speller grid, one of which is assumed to be the user's desired or target character during a selection process. To select a new target character, the user focuses on that character as subsets of characters are illuminated on the screen. The BCI infers the user's intended character by: i) processing EEG data following each stimulus presentation; ii) using a classifier to detect P300 event related potentials (ERPs) embedded in EEG data that are elicited in response to presentation of the target character; and iii) estimating the user's intended character by matching the character presentation patterns to the detected ERPs with a decoding algorithm.

In general, the experiment session consists of a calibration phase and a test phase. During the calibration phase, participants perform copy-spelling with no BCI feedback to collect labeled EEG data to train a P300 classifier. The amount of training data for each study is the same across all participants. During the test phase, the trained P300 classifier was applied and participants performed copy-spelling with the BCI prediction presented as feedback to evaluate a new BCI algorithm or strategy. The amount of test data within a study is either the same across all participants or varies across participants for studies with a dynamic stopping criterion for data collection.

3.3 Data Content

We extracted the relevant biosignal data from the BCI 2000 *.dat* files, enriched with additional data elements such as participant demographics (when available) and study information, and reformatted to the EDF+(24). EDF is an open format for multi-channel biological and physical signals. We chose the enhanced format, EDF 'plus' (EDF+) because it has additional specifications that includes technical specifications and individual attributes; it is widely used to store clinical EEG data. EDF+ file readers are various software platforms, including MATLAB, R, Python and C++ (25).

We have defined a new data dictionary for the file header and data records in the EDF+ file that encompasses the BCI data levels outlined in the IEEE P2731 *Standard for a Unified Terminology for Brain-Computer Interfaces* for data storage and sharing (9):

***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 (26)), 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 feedback. In general, no feedback is presented during the calibration phase to obtain labeled data to train the classifier and feedback is presented during the test phase to evaluate the trained classifier. We distinguish between the BCI prediction and the presented BCI feedback because the latter can be different in certain experimental paradigms, e.g., the fake feedback paradigm described later in the paper.

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

(a) EDF file header.

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#>*
Level 2: BCI Feedback		Target character	CurrentTarget
		Predicted target character	Selected<Target/Row/Column>
		Presented BCI feedback	DisplayResult, FakeFeedback

(b) EDF data

Table 2: Our data dictionary for the BCI file header and data records in EDF+ files based on developing IEEE P2731 standards for BCI data storage and sharing. 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.

The EDF+ file header contains a *patient* identification (ID) field with sub-fields to include information about individual attributes and a *recording* ID field with sub-fields to include information related to recording setting. Our BCI dataset does not contain personal identifiers. Instead, we modified the patient ID sub-fields to include participant demographics and modified the record ID sub-fields to include the study identifier and experiment session number, Table 2a. BCI data organized by the proposed IEEE P2731 BCI data levels and the data labels are shown in Table 2b. Most of the EDF+ data labels we used are derived from parameter definitions in BCI2000 .dat files (10).

The current batch of our BCI dataset (Batch 1) contains data from 11 previous BCI studies with a total of 159 participants, including 29 participants with ALS. The rest of the dataset (Batch 2) is currently undergoing data curation and engineering and contains data from 7 studies with up to 148 participants, including 20 participants with ALS. The data analysis focuses on Batch 1 of our BCI dataset. A summary of the study characteristics are summarized in Table S**??**.

4 Data Analysis

One of the unique aspects of our BCI data format is the inclusion of several BCI data attributes that facilitate reusability (27). This data analysis highlights the *multi-purpose* use of our BCI dataset: it can be used for analyses related to P300 ERP and BCI feedback responses.

4.1 Methods

4.1.1 Signal preprocessing

For each channel, a time window of EEG data was extracted from the task-relevant onset, which is either the illumination of characters on the grid or the presentation of BCI feedback. 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) (28) or a 16-channel subset (F3, Fz, F4, T7, C3, Cz, C4, T8, CP3, CP4, P3, Pz, P4, PO7, PO8, Oz).

Data resampling. This included block-averaging of non-overlapping time segments within the EEG time window (28) 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 (29).

After preprocessing, channel-specific features were concatenated to obtain a feature vector.

4.1.2 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 (30) and is still popularly used, including in all our BCI studies. The stepwise LDA is the default classifier in BCI2000 (31).

Convolutional neural network (CNN). EEGNet is a compact CNN architecture that was developed for generic classification in EEG-based BCIs (2). 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 (6). EEGNet consist 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. (32) 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 (32).

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).

4.2 Tasks

4.2.1 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 subset and raw EEG data with minimal preprocessing are typically used as input features to deep learning models (3; 4; 5); this is based on the rationale of leveraging model complexity to automatically learn robust feature representations from raw data without the need for manual feature extraction. However, the high data dimensionality has implications on model performance given the typical amounts of available BCI training data relative to the number of trainable parameters in the 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., (2), while the amount of user-specific data ranges from 2000-4000 observations for training the P300 classifier. We conducted experiments to investigate the impact, if any, of various signal preprocessing on the performance of EEG and CNN-LSTM models for P300 classification.

Raw features were extracted from an time window of 0.625 seconds. The following preprocessing steps were applied: channel subset (8, 16), downsampling decimation factor (none, 4, 8) and spatial filtering (with xDAWN, without xDAWN). For this analysis, we used the LDA model in the *scikit-learn* package, the EEGNet,8-2 model provided by the model developers via GitHub (2), and our implementation of the CNN-LSTM based on the specification in (32), 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 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.

4.2.2 Towards Generalized Error Detection in BCI Spellers

Error-related potentials (ErrPs) are neural signal components that are elicited when a person perceives erroneous actions or behavior. ErrPs have been proposed as an endogenous means to assess the accuracy of BCI predictions as it is assumed that an ErrP is elicited when the user observes erroneous BCI feedback. However, there has been limited use of ErrPs within the context of P300-based BCIs due to low accuracy with single trial detection and the time constraints involved with obtaining sufficient training data (33). For example, consider a typical trial of a P300-based BCI that yields 120 observations per character (corresponding to the number of sensory stimuli presented) to train a P300 classifier and only one observation per character (after BCI feedback presentation) to train an ErrP classifier. Moreover, for BCI users with high accuracy, the error class could be rare or nonexistent.

To ensure enough data and samples of the error class, a separate calibration phase is typically conducted to train 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. One of the studies in our dataset, which we denote the error-related negativity (ERN) study (Study C in Table **??**) includes a conventional calibration phase to train a P300 classifier and a fake feedback phase to train an ErrP classifier. However, a separate ErrP classifier calibration phase is 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 when applied to

novel users without retraining. This included two approaches: using an ensemble of subject-specific classifiers or a generic classifier trained on data pooled from various BCI users.

Data from the fake feedback phase of each participant in the ERN study consists of 300 observations with an error rate of 20%. ErrP classifier models were trained on data from subjects in the ERN study (source subjects) and applied to data from other subjects (target subjects) from either the same study or other studies. First, we investigated the performance of LDA and EEGNet 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 (2). To generate the classifier ensemble, user-specific LDA classifiers were trained using leave-one-word-out cross-validation. For the generic classifier, EEGNet models were trained on data pooled across subjects 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 two EEG window lengths, short (1.25 or 1.3 secs) and long (2 secs), for all classifiers and two feature types for the LDA classifier (block-averaged and xDAWN features enhanced with Riemannian geometry (RG) manifold projection (29; 34).

The performance of the trained ErrP classifiers on error detection in the P300 speller was evaluated. For generic ErrP classification for the other studies, data from all subjects in the ERN study were pooled to train an EEGNet classifier. The ensemble-based classification decision was based on a voting scheme, with each source subject's classifier's vote weight assigned either uniformly or based on the similarity between the source and target subjects' 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 target subjects' EEG data and normalized across source subjects (35).

4.2.3 Statistical Analyses

Our statistical analyses account for repeated measures from each subject 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 model, with Aikake information criterion (36) values to compare potential models. Character accuracy was the response variable, with subject as a random effect, and the following fixed effects: model, application of spatial filtering method, number of EEG channels and downsampling decimation factor, and the interaction effects between model and each of the other fixed effects. Then, performance was visualized using estimated marginal means interaction plots (*emmeans* package in R). The statistical significance of parameter coefficients was assessed using t-tests with Satterwaithe's method, $p < 0.05$. For the ErrP experiments, we assessed statistical significance using the Wilcoxon signed-rank tests ($p < 0.05$) to compare the effect of EEG window length and classifier model on AUCs, with Bonferroni adjustments for pairwise comparisons.

5 Results

5.1 Demographics

Figure 1 summarizes population-specific demographics, including the proportion of missing data, extracted from the EDF+ file metadata of the current version of our current BCI dataset. Most of the missing demographic data are from our earlier studies where the data were either not adequately preserved or not obtained from participants. 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. We primarily recruit from a university population for practical reasons to conduct several preliminary online studies with fewer logistical hurdles prior to testing promising candidates in the ALS population.

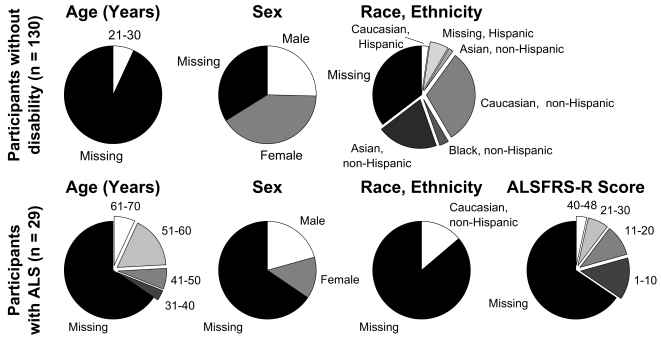

Figure 1: Summary of participant demographics of the current batch of our BCI dataset extracted from the file metadata. ALSFRS-r, revised amyotrophic lateral sclerosis (ALS) Functional Rating Scale.

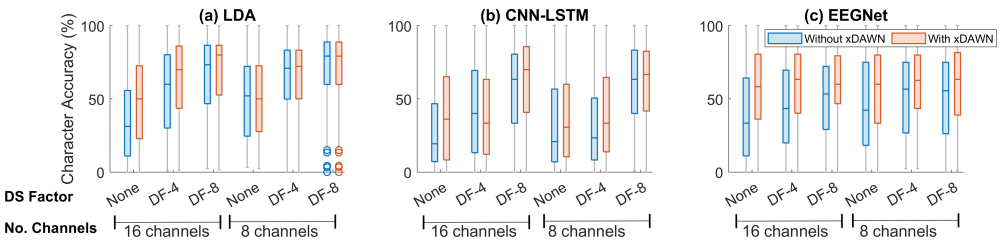

Figure 2: Box plots of P300 speller character accuracy with conventional and deep learning models, given various signal preprocessing methods. (a) Conventional LDA classifier; (b) CNN-LSTM (32); (c) EEGNet (2). Results are grouped by downsampling decimation factor (DF) and channel subset.

5.2 P300 Analysis

Character selection accuracy with subject-specific P300 classifiers across subjects are summarized in Figure 2 using boxplots to highlight the performance range across subjects. (Subject-specific performances across the 36 conditions, estimated marginal means interaction plots and summaries of statistical tests are included in the supplement.) 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 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.

5.3 ErrP Analysis

Cross-subject performance of the generic EEGNet ErrP classifiers (raw features, short vs. long window) and the within-subject performance of LDA classifiers (block averaged vs. xDAWN features, short vs. long window) are shown in Figure 3. Statistical results are summarized in Table **??**. For each model and input feature combination, features extracted from the long EEG windows achieved statistically significantly better performance relative to the short window, $p < 0.05$. The rest of the analysis uses the best performing feature set for each model: (LDA, xDAWN-RG features) and (EEGNet, raw features) both extracted from a 2-second window.

Figure 4 shows the precision and recall of error detection with the ensemble-based LDA and generic EEGNet ErrP classifiers. The LDA classifier ensemble with uniform weighting is not useful as it almost always predictions a single class. While the LDA classifier ensemble with similarity-based

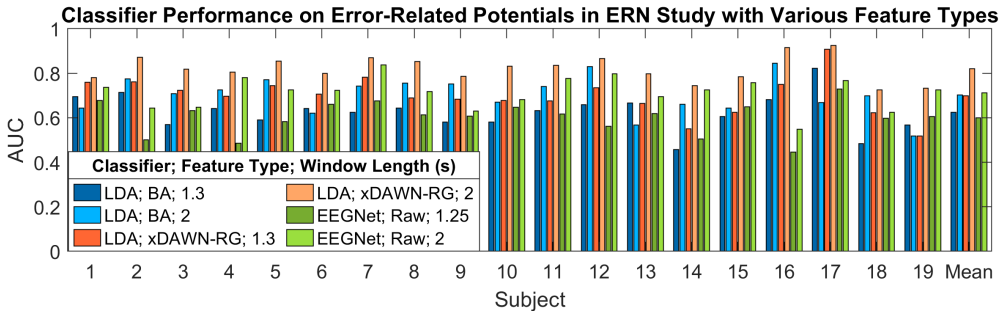

Figure 3: Areas under the receiver operating characteristics curve (AUC) for various error-related potential classifier models trained on data from the error-related negativity (ERN) study. Subject-specific linear discriminant analysis (LDA) classifiers were trained with leave-one-word-out on block-averaged (BA, blue) or xDAWN-Riemannian geometry (RG) features (orange) extracted from 1.3s or 2.0s time windows. Generic EEGNet classifiers were trained on raw features extracted from either 1.25s and 2.0s windows with leave-one-subject out folds.

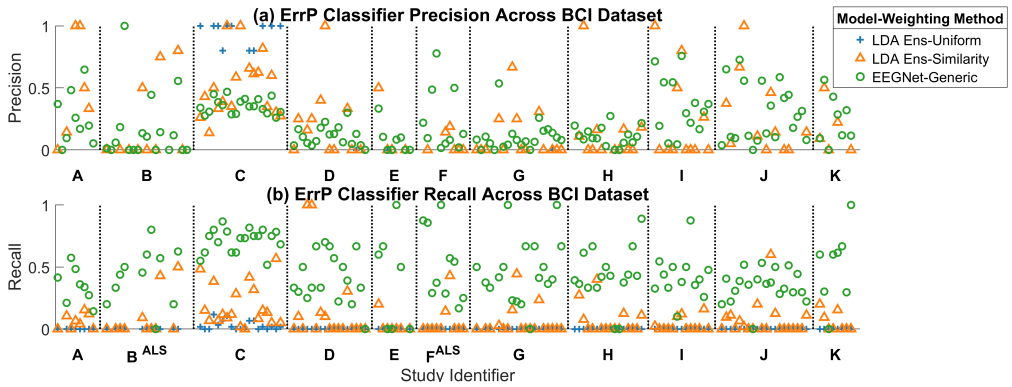

Figure 4: (a) Precision and (b) recall of error detection with error-related potential (ErrP) classifiers trained on data from the error-related negativity (ERN) study (Study C) and applied to target subjects segregated by study. Classifier models included an ensemble of subject-specific LDA classifiers 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.

weighting generated more varied predictions, it performance, particular recall, was still poor. EEGNet had the highest recall performance. Overall, the results highlight challenges with single trial error detection with ErrPs.

6 Discussion and Future Work

Our analyses presented here are not meant to be conclusive or comprehensive as primary goals were to showcase the broad performance trends, test data quality and demonstrate reusability in order to refine our data engineering protocol to a near-final version; thus, we limited the number of conditions for iterative 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 us to run a wide range of experiment where subsets of data can be carved out if similar conditions are needed across participants.

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 review. We also highlight several lessons, which are likely not unique to our dataset development process. Our BCI data spans over a decade of BCI research. This extended time period creates challenges beyond missing data, including different BCI2000 implementations and varied data archiving format across experimenters over the years, all of which hinder automated data processing and in house data consolidation. These issues have been mitigated in recent years as we have standardized data collection instruments and migrated from paper to electronic forms.

Our proposed large, diverse, machine learning ready BCI dataset provides a new resource to facilitate robust and well-powered analyses for BCI research. To the best of our knowledge, our final BCI dataset would represent the largest collection of P300 speller data from individuals with ALS ($N = 29$ in the current batch, $N = 49$ in total) in a single dataset. We plan to publicly release our final BCI dataset under a creative commons license via the open PhysioNet repository. It is possible that our meta and data definitions may need to be modified once the IEEE P2731 standards for BCI data storage and sharing are finalized. As we have defined our BCI data format based on the proposed BCI data levels and optimized our data generation protocol, we can easily adapt our BCI dataset to be compliant with future official BCI data standards.

Acknowledgments and Disclosure of Funding

This work was funded by a US Federal Agency.

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] 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," *J. Neural Eng.*, vol. 15, no. 5, p. 056013, 2018.

[3] 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," *J. Neural Eng.*, vol. 15, no. 3, p. 031005, 2018.

[4] 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.

[5] 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.

[6] 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.

[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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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 Trans Biomed Eng*, vol. 51, no. 6, pp. 1044–51, 2004.

[14] 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 Trans Neural Syst Rehabil Eng*, vol. 14, no. 2, pp. 153–9, 2006.

[15] 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)?" *Neurosci Lett*, vol. 462, no. 1, pp. 94–8, 2009.

[16] L. Citi, R. Poli, and C. Cinel, "Documenting, modelling and exploiting p300 amplitude changes due to variable target delays in donchin's speller," *J Neural Eng*, vol. 7, no. 5, p. 056006, 2010.

[17] M. S. Treder, N. M. Schmidt, and B. Blankertz, "Gaze-independent brain-computer interfaces based on covert attention and feature attention," *J Neural Eng*, vol. 8, no. 6, p. 066003, 2011.

[18] 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.

[19] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245713000606

[20] 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," *Front Hum Neurosci*, vol. 7, p. 732, 2013.

[21] 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. [Online]. Available: https://doi.org/10.1038/s41597-022-01509-w

[22] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B. A. S. Group, and A. c. l. o. t. B. S. Group, "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.

[23] "BCI2000 Contributions: Eyetrackerlogger." [Online]. Available: https://www.bci2000.org/mediawiki/index.php/Contributions:EyetrackerLogger

[24] 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.

[25] [Online]. Available: https://www.edfplus.info/

[26] "Racial and ethnic categories and definitions for nih diversity programs and for other reporting purposes, notice number: Not-od-15-089," 2015.

[27] 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.

[28] 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," *J Neural Eng*, vol. 3, no. 4, pp. 299–305, 2006.

[29] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xdawn algorithm to enhance evoked potentials: application to brain-computer interface," *IEEE Trans Biomed Eng*, vol. 56, no. 8, pp. 2035–43, 2009.

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

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

[32] 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.

[33] M. Yasemin, A. Cruz, U. J. Nunes, and G. Pires, "Single trial detection of error-related potentials in brain-machine interfaces: A survey and comparison of methods," *Journal of Neural Engineering*, 2022.

[34] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain-computer interface classification by riemannian geometry," *IEEE Trans Biomed Eng*, vol. 59, no. 4, pp. 920–8, 2012.

[35] 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.

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

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

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 Section 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? **[TODO]**
   (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 4.2.3 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 4.1.2.
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? [No] The dataset includes data from data from several studies. A general description of the BCI protocol are provided in Section 3.2.3
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] The dataset includes data from data from several studies.
   (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

A Appendix

Include extra information in the appendix. This section will often be part of the supplemental material. Please see the call on the NeurIPS website for links to additional guides on dataset publication.

1. Submission introducing new datasets must include the following in the supplementary materials:
   (a) Dataset documentation and intended uses. Recommended documentation frameworks include datasheets for datasets, dataset nutrition labels, data statements for NLP, and accountability frameworks.
   (b) URL to website/platform where the dataset/benchmark can be viewed and downloaded by the reviewers.
   (c) Author statement that they bear all responsibility in case of violation of rights, etc., and confirmation of the data license.
   (d) Hosting, licensing, and maintenance plan. The choice of hosting platform is yours, as long as you ensure access to the data (possibly through a curated interface) and will provide the necessary maintenance.
2. To ensure accessibility, the supplementary materials for datasets must include the following:
   (a) Links to access the dataset and its metadata. This can be hidden upon submission if the dataset is not yet publicly available but must be added in the camera-ready version. In select cases, e.g when the data can only be released at a later date, this can be added afterward. Simulation environments should link to (open source) code repositories.
   (b) The dataset itself should ideally use an open and widely used data format. Provide a detailed explanation on how the dataset can be read. For simulation environments, use existing frameworks or explain how they can be used.
   (c) Long-term preservation: It must be clear that the dataset will be available for a long time, either by uploading to a data repository or by explaining how the authors themselves will ensure this.
   (d) Explicit license: Authors must choose a license, ideally a CC license for datasets, or an open source license for code (e.g. RL environments).
   (e) Add structured metadata to a dataset's meta-data page using Web standards (like schema.org and DCAT): This allows it to be discovered and organized by anyone. If you use an existing data repository, this is often done automatically.
   (f) Highly recommended: a persistent dereferenceable identifier (e.g. a DOI minted by a data repository or a prefix on identifiers.org) for datasets, or a code repository (e.g. GitHub, GitLab,...) for code. If this is not possible or useful, please explain why.
3. For benchmarks, the supplementary materials must ensure that all results are easily reproducible. Where possible, use a reproducibility framework such as the ML reproducibility checklist, or otherwise guarantee that all results can be easily reproduced, i.e. all necessary datasets, code, and evaluation procedures must be accessible and documented.
4. For papers introducing best practices in creating or curating datasets and benchmarks, the above supplementary materials are not required.