A Data-Centric Analysis of the Impact of Training Data Quality vs. Quantity on P300 Brain-Computer Interface Performance (Student Abstract)

Arnav Gupta[1*], Albert Liu[1*], Eliza Haines[2*], Riyadh Alghamdi[3*], Aniketh S. Kota[1], Leslie M. Collins[1], Boyla O. Mainsah[1†]

[1]Duke University, Durham, NC, USA
[2]Clemson University, Clemson, SC, USA
[3]King Abdullah University of Science and Technology, Thuwal, KSA
boyla.mainsah@duke.edu

Abstract

The current standard for training brain-computer interface (BCI) machine learning models is user-specific. There is a high interest in developing generic models that are trained on data from other users to minimize BCI calibration time; however, this is limited by noisy, non-stationary brain signals and high inter-user variability. We investigate the trade-off between training data quality and quantity on P300 BCI performance in individuals with amyotrophic lateral sclerosis (ALS) with representative traditional machine learning (stepwise linear discriminant analysis, SWLDA) and deep learning (EEGNet) models. Results show that data quality and domain alignment are more critical than dataset size: user-specific models trained on significantly less data outperformed generic models; generic models trained on ALS data outperformed models trained on non-ALS data; block-averaging of features was mostly detrimental to EEGNet but beneficial to SWLDA; and accounting for inter-stimulus interval differences between ALS and non-ALS data had minimal effect. Our findings highlight the importance of individualized model tuning for reliable P300 BCIs.

Introduction

The P300-based BCI (Farwell and Donchin 1988) is a widely researched BCI in individuals with severe neuromuscular limitations, such as late stage amyotrophic lateral sclerosis (ALS). A main limitation with electroencephalography (EEG)-based BCIs is high within-user and inter-user signal variability due to noisy, non-stationary EEG data. Thus, the current conventional approach to train machine learning models in BCI applications is user-specific. Alternatively, the development of generic models trained on data from other users is a promising way to reduce the burden of user-specific calibration in target BCI end-user populations.

However, challenges to transfer learning still remain: high variability in EEG signals across individuals, differences in experimental parameters across BCI populations and limited publicly available data from target end-user populations. For generic models, there is also the need to balance model size against the amount of training data available and time constraints for real-time BCI signal processing. Recent studies indicate that the current paradigm of scaling model size to increase performance is not necessarily as effective in EEG-based BCIs. In some BCI applications, smaller convolutional neural network (CNN) models (e.g., EEGNet) achieve comparable performance as large EEG foundation models (Lee et al. 2025), and CNN-only encoder models perform comparably or better than CNN + transformer counterparts (Kostas, Aroca-Ouellette, and Rudzicz 2021; Cui et al. 2024; Lee et al. 2025).

Limits of a model-centric approach with noisy, non-stationary EEG data suggest a data-centric approach for developing generic models for BCI applications. In this work, we investigate the trade-offs between training data quality and quantity in a target BCI end-user population with the P300-based BCI.

Methodology

Dataset. We used the bigP3BCI dataset (Mainsah et al. 2025) containing data obtained from prior P300 BCI speller studies with ALS and non-ALS participants. ALS studies used an inter-stimulus interval (ISI) of 250 ms and checkerboard (CB) paradigm with a speller grid size of 6×6 while non-ALS studies used an ISI of 125 ms and CB paradigm with a 9×8 grid. The data domains include: ***user-specific,*** training data from a participant in the held-out ALS study (Study L); ***within-domain*** (WD), training data from other ALS participants (Study B); and ***cross-domain*** (CD), training data from non-ALS participants (Studies A, C, H, K, M).

Preprocessing. EEG signals from electrodes [Fz, Cz, P3, Pz, P4, PO7, PO8] at 256 Hz sampling rate were notch (60 Hz) and bandpass (0.5-30 Hz) filtered and segmented into stimulus-locked epochs (0–800 ms). Features used included either the full filtered EEG signals (8 channels × 206 samples and an added convolution dimension for EEGNet models) or block-averaged features at a rate of 20 Hz. To account for ISI differences between non-ALS (125 ms) and ALS (250 ms) studies, non-ALS data with an assumed ISI of 250 ms was also used (referred to as *ISI matching*).

*These authors contributed equally.
†Corresponding author.

P300 Classification. Models were trained to distinguish between EEG responses to target (P300 class) and non-target stimuli, with increasing amounts of training data. Data were shuffled once before training with 20% of training trials reserved for validation. Classifier models evaluated included:

- ***Stepwise linear discriminant analysis*** (SWLDA), the most widely used traditional machine learning model (Farwell and Donchin 1988; Kalra et al. 2023). Features were z-scored. SWLDA significance thresholds were set to p_{enter} = 0.1, p_{remove} = 0.15, and max_{iter} = 60.

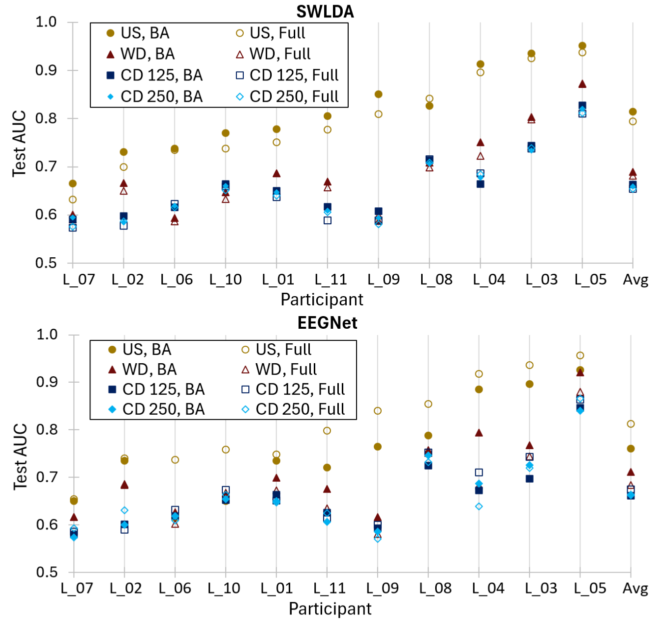
- ***EEGNet***, a compact CNN for real-time feasibility (Lawhern et al. 2018), which has been tested in online BCI studies (Lee et al. 2020). Hyperparameters were set to: F1 = 8 temporal filters; D = 2 depthwise filters per F2 = 16 or 206 pointwise filters for sampling rate of 20 Hz or 256 Hz, respectively; kernel length = 16; and dropout = 0.5. A reduced sampling rate yields a smaller EEGNet model. The number of trainable parameters was 818 and 3778 for sampling rates of 20 Hz (with block-averaging) and 256 Hz, respectively. Models were trained for 200 epochs with Adam optimizer, a learning rate of 0.0001 and categorical cross-entropy loss weighted by the class frequency. The checkpoint with lowest validation loss was used.

Results

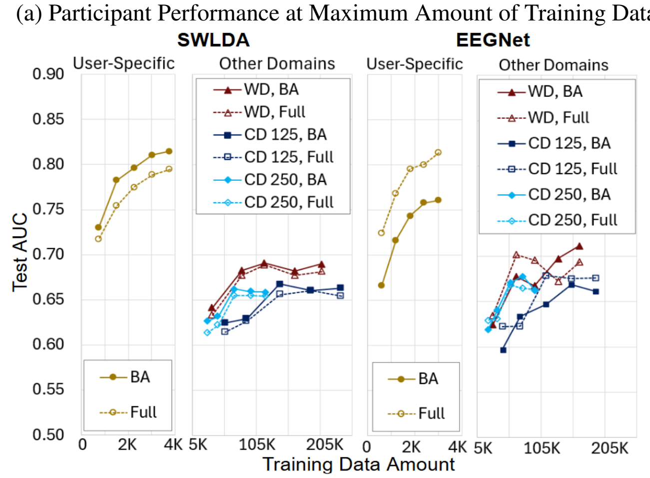
P300 classifier performance was evaluated using the area under the receiver operating characteristic curve (AUC). A linear mixed effects model was fit with fixed effects of classifier model, data resampling and training data size and random effect of training data size grouped by participant.

Figure 1a shows test AUC values of participants in Study L of the bigP3BCI dataset at the maximum amount of training data for each data domain condition. Figure 1b shows mean test AUC values as a function of the amount of training data. Across all conditions, performance improved with additional training data, but gains diminished beyond a certain threshold. Overall, user-specific models consistently yielded the highest performance and WD models outperformed CD models for both SWLDA and EEGNet. The performance gap between user-specific and generic models widened as user-specific test AUC performance increased, Figure 1a.

For SWLDA and EEGNet, ISI matching the CD data from 125 ms to 250 ms had minimal impact on performance. While block-averaging of features consistently improved the performance of SWLDA models, it mostly decreased the performance of EEGNet models.



(a) Participant Performance at Maximum Amount of Training Data.



(b) Average (Avg) Performance vs. Amount of Training Data.

Figure 1: Area under the receiver operating characteristic curve (AUC) of user-specific (US), within-domain (WD) and cross-domain (CD) P300 classifiers (stepwise linear discriminant analysis, SWLDA; EEGNet). Features include the full filtered signals from stimulus-locked epochs (800 ms) at 256 Hz or block-averaged (BA) features at 20 Hz. CD model training used either the default inter-stimulus interval (ISI) of 125 ms (CD 125) or ISI matching at 250 ms (CD 250).

Discussion

The general performance trends indicate that data quality and data domain alignment are more important than dataset size. Limitations include the small sample size of the test ALS study, though a broad range of participant performance levels are represented. User-specific models trained on very limited data significantly outperformed generic models trained on larger but less aligned data, underscoring the importance of individualized models. The impact of block-averaging was model dependent: it was generally beneficial to SWLDA but detrimental to EEGNet. ISI matching, which reduced the non-ALS training data by about 50%, did not significantly affect performance; this suggests that ISI mismatch is less critical than data domain mismatch.

Future work should investigate advanced feature preprocessing and domain adaptation methods that learn robust task-specific EEG data representations. Emphasizing data-centric machine learning with user-specific fine tuning will be essential for advancing P300 BCIs toward reliable use in target BCI end-user populations.

Acknowledgments

This work was funded by the 2025 Data+ Summer Research Program at Duke University.

References

Cui, W.; Jeong, W.; Thölke, P.; Medani, T.; Jerbi, K.; Joshi, A. A.; and Leahy, R. M. 2024. Neuro-GPT: Towards a foundation model for EEG. In *IEEE International Symposium on Biomedical Imaging (ISBI)*.

Farwell, L. A.; and Donchin, E. 1988. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6): 510–523.

Kalra, J.; Mittal, P.; Mittal, N.; Arora, A.; Tewari, U.; Chharia, A.; Upadhyay, R.; Kumar, V.; and Longo, L. 2023. 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*, 31: 1429–1439.

Kostas, D.; Aroca-Ouellette, S.; and Rudzicz, F. 2021. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15: 653659.

Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5): 056013.

Lee, J.; Won, K.; Kwon, M.; Jun, S. C.; and Ahn, M. 2020. CNN with large data achieves true zero-training in online P300 brain-computer interface. *IEEE Access*, 8: 74385–74400.

Lee, N.; Barmpas, K.; Panagakis, Y.; Adamos, D.; Laskaris, N.; and Zafeiriou, S. 2025. Are Large Brainwave Foundation Models Capable Yet? Insights from Fine-Tuning. In *International Conference on Machine Learning (ICML)*.

Mainsah, B.; Fleeting, C.; Balmat, T.; Sellers, E.; and Collins, L. 2025. bigP3BCI: An Open, Diverse and Machine Learning Ready P300-based Brain-Computer Interface Dataset (version 1.0.0). PhysioNet, RRID:SCR_007345.