

Evaluating the Impact of Generic Model Priors on User-Specific Calibration in P300 Brain-Computer Interfaces

Arnav Gupta, Boyla O. Mainsah, Leslie M. Collins

Department of Electrical and Computer Engineering, Duke University, Durham, NC, United States

E-mail: {arnav.gupta, boyla.mainsah, leslie.collins}@duke.edu

The current standard approach for training brain-computer interface (BCI) models is user-specific. Due to high inter-individual variability, there is often a performance gap between generic BCI models trained on data from other users relative to user-specific models. Alternatively, generic models can provide informative priors to reduce BCI calibration time. We investigate the impact of model priors on shrinkage linear discriminant analysis (LDA) with Bayesian sequential updates using a normal-inverse-Wishart (NIW) conjugate prior. Baseline WD and CD P300 classifiers achieved off-the-shelf performance of 0.629 ± 0.022 and 0.612 ± 0.020 area under the curve (AUC), respectively. Incorporating 20% user-specific calibration data (total = 3780 samples) with WD and CD priors yielded a statistically significant increase to 0.683 ± 0.022 AUC with convergence of both models to user-specific performance 0.699 ± 0.023 . Sequential updates of shrinkage LDA (20% data increments) required on average 0.001 seconds with WD/CD priors vs. 0.140 seconds with only user-specific data, demonstrating that generic informative priors efficiently handle domain shifts with reduced computational overhead.

INTRODUCTION

The P300 brain-computer interface (BCI) speller [1] is a widely studied BCI paradigm that provides a means of communication for individuals with severe neuromuscular impairments, such as amyotrophic lateral sclerosis (ALS). In a typical visual P300 BCI speller, items are arranged in a grid interface and rows and columns are sequentially illuminated in a randomized sequence (Fig. 1). When the user's desired target character is illuminated - a relatively rare event - a P300 event related potential (ERP) response is elicited approximately 300 ms after the stimulus is presented. By analyzing segments of electroencephalography (EEG) signals synchronized with stimulus onset, the BCI identifies the user's intended character with a character decoding function (Fig. 2).

However, EEG signals are noisy and non-stationary, making reliable P300 classification challenging. Typically, user-specific BCI decoding models are trained on labeled data collected from the target user during a calibration session. While generic models trained on data from other users offer a potential to reduce BCI calibration time, their performance is often inferior to that of



Figure 1: (Left) Individual using a P300 speller. (Right) BCI spelling interface [2].



Figure 2: (Left) Average electroencephalography (EEG) signals from participant L_03 following target and non-target stimuli. A characteristic positive peak at approximately 270 ms is observed for target stimuli relative to non-target stimuli. (Right) Heatmap of character scores following cumulative P300 classifier score updates; character "V" receives the highest score and is selected as the user's intended character [2].

user-specific models due to inter-user variability and differences in population data domains [3]. Prior work has shown that data domain alignment plays a critical role in model performance, with user-specific P300 classifier models trained on limited data often outperforming generic models trained on larger but less aligned data [2]. The goal of this study is to bridge the gap between calibration-heavy user-specific models and lower-performing generic models in P300 BCIs with transfer learning. This work investigates if the use of informative model priors can mitigate the performance gap with generic models while reducing user BCI calibration time. Integrating incremental amounts of target user data into generic baselines, we quantify the amount of calibration data needed to achieve user-specific performance.

MATERIALS AND METHODS

Dataset: We used data from the bigP3BCI dataset, which contains EEG recordings from previous P300

speller studies with participants with and without ALS [4]. The ALS studies utilized a 6×6 speller grid with a 250 ms inter-stimulus interval (ISI), whereas the non-ALS studies utilized a 9×8 grid with a 125 ms ISI. For consistency, we limited to experiments with the checkerboard stimulus presentation paradigm [5].

Data Domains: Tab. 1 summarizes the study characteristics and the corresponding domain assignments. Population domain definitions include:

- User-specific (US): Training data from the target participant in the held-out test ALS study (Study L).
- Within-domain (WD): Training data from participants with ALS (Study B).
- Cross-domain (CD): Training data from participants without ALS (Studies A and C).

Table 1: Overview of P300 BCI studies used from the bigP3BCI dataset and domain definitions. ALS, amyotrophic lateral sclerosis; CD, cross-domain; US, user-specific; ISI, inter-stimulus interval; WD, within-domain.

Study	ISI (ms)	Grid Size	Population	Domain
A	125	9×8	Non-ALS	CD
B	250	6×6	ALS	WD
C	125	9×8	Non-ALS	CD
L	250	6×6	ALS	US

For baseline WD and CD configurations, a subset of 80,000 epochs was randomly drawn from each respective domain pool to train generic models; results were averaged from generic model training after three independent trials to minimize sampling variance. To account for the discrepancy in ISI between ALS (250 ms) and non-ALS (125 ms), we performed *ISI matching* [2] in the non-ALS data by retaining data from every other stimulus prior to generating the sampled subset. ISI matching is implemented on the stored data and does not alter the original recordings or the feedback delivered during the non-ALS sessions.

For the held-out study, user-specific data contained 3,780 user epochs each for training and testing. User-specific model parameters were updated after adding data sequentially in per-word increments of 756 epochs (20% per word increment), until the maximum amount of training data available.

Preprocessing: EEG signals from electrodes Fz, Cz, P3, Pz, P4, PO7, PO8, Oz were used for signal analysis [6]. Signals were downsampled from 256 Hz using block averaging at 20 Hz. A notch filter was applied at 60 Hz to suppress power line noise, followed by a band-pass filter (0.5–30 Hz) to isolate the P300-relevant frequency band while attenuating slow drifts. No automated artifact rejection was applied. Stimulus-locked epochs were extracted from 0–800 ms following each stimulus onset. Data across channels in epochs were vectorized into a $1 \times D$ feature vector prior to classifier training.

Sequential Classifier Updates: We selected a linear discriminant analysis (LDA) as a classifier model that is amenable to sequential updates. Features are assumed to be multivariate Gaussian variables. The LDA weight vector is defined as:

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \quad (1)$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ are mean vectors of target and non-target classes, respectively; $\mathbf{\Sigma} = p_1\mathbf{\Sigma}_1 + p_0\mathbf{\Sigma}_0$ is the shared covariance matrix; and p_c is the class-specific proportion.

Shrinkage linear discriminant analysis (LDA) was first proposed for P300 classification in [7] to address the ill-conditioned nature of empirical covariance matrices in high-dimensional EEG feature spaces. In shrinkage LDA, the empirical covariance matrix, $\mathbf{\Sigma}$, towards a structured target matrix $\mathbf{T}$ via a shrinkage intensity coefficient $\lambda \in [0, 1]$:

$$\mathbf{\Sigma}_{\text{shrink}} = (1 - \lambda)\mathbf{\Sigma} + \lambda\mathbf{T} \quad (2)$$

In this work, we will update shrinkage LDA weights sequentially via Bayesian inference with a normal-inverse-Wishart (NIW) conjugate prior over the class-specific means and shared covariance [8]. The initial baseline class-specific parameters (mean vector $\mathbf{m}_0$, scaling factor κ_0 , degrees of freedom ν_0 , and scatter matrix derived from the shared covariance, $\mathbf{S}_0 = \nu_0 * \mathbf{\Sigma}_0$, define the prior model. Given an incoming block of data $\mathbf{X} \in \mathbb{R}^{D \times N}$ containing N samples with D -dimensions with a sample mean of $\bar{\mathbf{x}}$, the NIW distribution parameters are updated accordingly:

$$\kappa_N = \kappa_0 + N \quad (3)$$

$$\nu_N = \nu_0 + N \quad (4)$$

$$\mathbf{m}_N = \frac{\kappa_0\mathbf{m}_0 + N\bar{\mathbf{x}}}{\kappa_N} \quad (5)$$

$$\mathbf{S}_N = \mathbf{S}_0 + \mathbf{X}\mathbf{X}^T + \kappa_0\mathbf{m}_0\mathbf{m}_0^T - \kappa_N\mathbf{m}_N\mathbf{m}_N^T \quad (6)$$

The initial class-specific mean and covariance priors were set based on within-domain, cross-domain or user-specific priors (using *scikit-learn* package). κ_0 defines the weight of the initial mean prior while ν_0 defines the weight of the initial covariance prior; we set $\kappa_0 = \nu_0 = 1$. At each data increment step i with N_i samples, shrinkage LDA parameters are updated with the maximum *a posteriori* (MAP) estimates of class-specific means and shared covariance:

$$\boldsymbol{\mu}_{i,c}^{\text{MAP}} = \mathbf{m}_{N_i,c} \quad (7)$$

$$\boldsymbol{\Sigma}_i^{\text{MAP}} = \frac{\mathbf{S}_{N_i}}{\nu_{N_i} + D + 1} \quad (8)$$

RESULTS

P300 Classifier Performance: Performance of trained P300 classifiers was evaluated on test data of participants

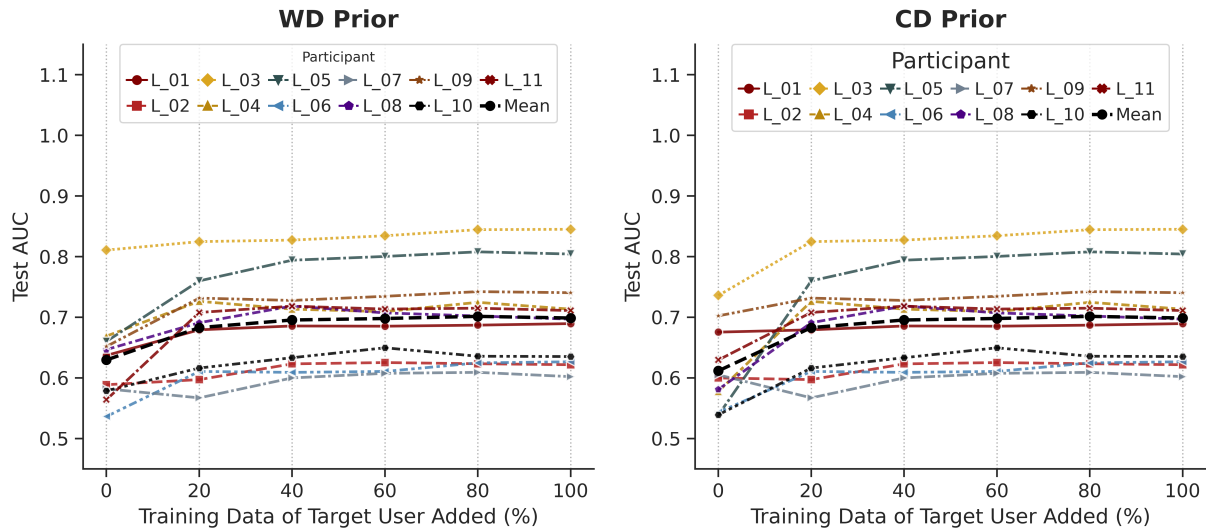


Figure 3: Test area under the curve (AUC) of shrinkage linear discriminant analysis (LDA) P300 classifiers with within-domain (WD) and cross-domain priors as a function of the amount of target user training data added (3780 total samples) for each participant.

in Study L in the bigP3BCI dataset. We analyzed how the incremental inclusion of target user data impacts classification performance across different training data domains. Fig. 3 shows the P300 classifier AUC results for WD and CD models, respectively as a function of the amount of training data from the target user for each participant. Fig. 4 shows the mean AUC values with sequential updates as a function of the amount of training data for generic and user-specific models. The generic WD baseline (0.629 ± 0.022) performs better than the generic CD baseline (0.612 ± 0.020), though this difference was not significant ($p = 0.37$). With incorporation of target user data, performance of both updated models with generic priors converge to that of user-specific models. Both generic models experience a significant initial performance gain with the first sequential update with 20% of target user data (0.683), moderate gains at 40% (0.695) and minimal gains until the full amount of calibration data (0.699).

To model progression of AUC performance in incoming sequential data, a linear mixed-effects (LME) model was fit with amount of calibration data and training data conditions as fixed effects and participant as a random intercept. The model revealed a highly significant main effect of amount of calibration percentage ($\beta = 0.001$, $p < 0.001$), confirming a reliable positive slope across calibration intervals. For trajectory variations from generic baselines, a second LME model restricted to the generic conditions was fit with an interaction term. The interaction between the amount of calibration data and generic training condition was non-significant ($\beta = 0.000$, $p = 0.421$, confirming that both the WD and CD priors follow a similar adaptation path when augmented with user-specific calibration data. Post-hoc paired tests between the initial generic baselines and counterpart sequentially updated models with the full calibration data revealed a highly significant performance gain (WD, $t = -5.286$,

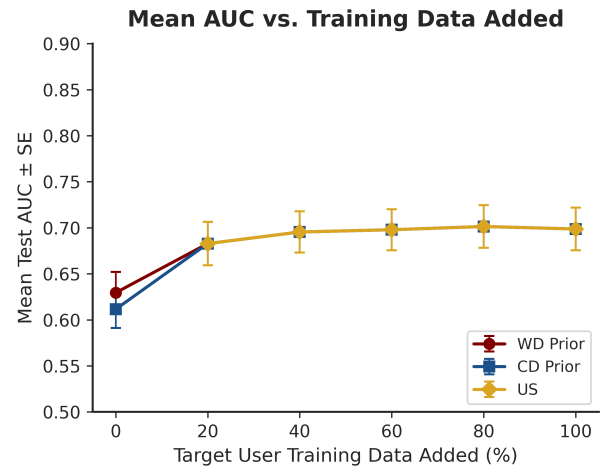


Figure 4: Mean test area under the curve (AUC) $\pm$ standard error (SE) of shrinkage linear discriminant analysis (LDA) P300 classifiers with within-domain (WD) or cross-domain prior and exclusively user-specific (US) data as a function of the amount of target user training data added (3780 total samples). Trajectories of mixed WD/CD converging towards US performance at 20% calibration.

$p = 0.0004$; CD, $t = -3.869$, $p = 0.0031$).

Fig. 5 shows the final classification performance across all conditions, tracking environments. At full calibration (100% added data), both sequentially update models with WD and CD priors achieved similar accuracy as user-specific models (0.699 ± 0.023). This confirms that regularized adaptation allows generic models to effectively overcome initial domain shifts.

Model Training Time: Fig. 6 shows the mean training time on our computing resource across training conditions. The baseline WD and CD models take a longer time due to a larger data pool; though this does not impact the training time of a novel user. Sequential updates with only user-specific data required 140 ms per data in-

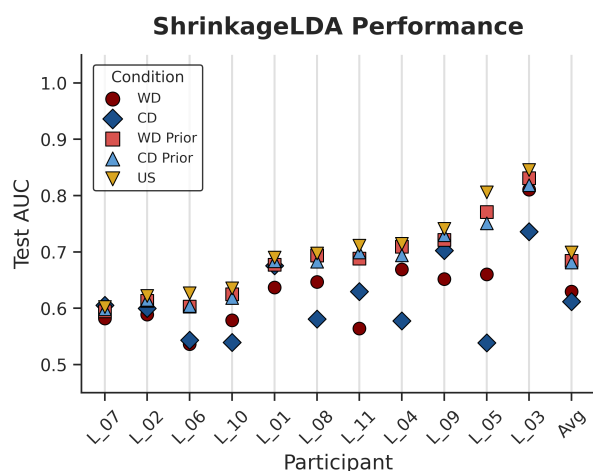


Figure 5: Area under the curve (AUC) of the receiver operating characteristic of shrinkage linear discriminant analysis (LDA) P300 classifiers with 100% calibration data. Markers denote participant test results for the various conditions and the mean performance (Avg). CD, cross-domain; US, user-specific; WD, within-domain.

crement while sequential updates with the WD and CD priors required 1 ms per increment. This represents an significant reduction in computational overhead with sequential updates vs. training on the final pooled data.

DISCUSSION

The analysis of shrinkage LDA with a Bayesian updating framework revealed a clear trend: when utilizing robust regularized linear classifiers, adaptive updates can effectively mitigate the impact of data domain mismatch. While within- or cross-domain data can introduce a persistent performance penalty, the integration of generic priors into local user parameters allows for rapid calibration and steady optimization of the decision space.

The Role of Sequential Calibration: Both within-domain and cross-domain shrinkage LDA models converged to to user-specific performance after a small initial amount of target-user data. This indicates that the NIW prior framework functions as an efficient regularizer. Instead of distorting the classification boundary, the historical prior data provides a robust foundation for the covariance structure, while the target user samples quickly align the class-specific means. This minimizes over-fitting to a limited amount of calibration data without retaining the unaligned structural biases of the baseline generic models.

Implications for BCI Deployment: While our bigP3BCI dataset includes data from individuals with ALS, data from target BCI end user populations are often difficult to obtain and not widely publicly available. In contrast, public BCI dataset mostly include data from individuals without neuromuscular limitations. Models with sequential updates using cross-domain priors achieved similar performance as that within-domain priors and user-specific models, demonstrating that domain

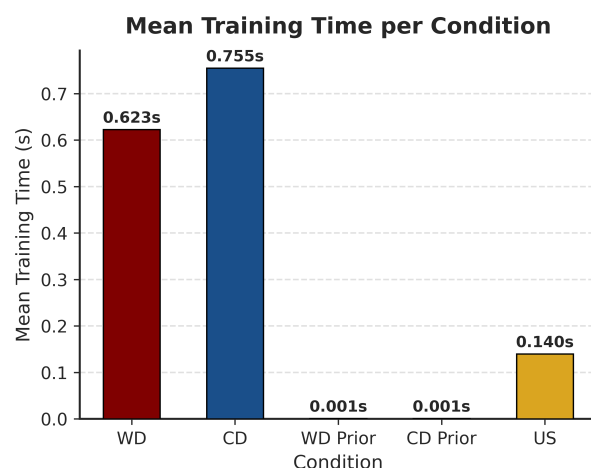


Figure 6: Mean training time of generic models and latency (seconds) of Bayesian sequential updates with generic prior (CD, cross-domain; WD, within-domain or only user-specific (US) data. The Xeon® Gold 6252 central processing unit (24-Core 2.1GHz) was used.

misalignment is not an absolute barrier. A BCI system can exploit generic model priors obtained from non target BCI end user populations, and potentially still achieve user-specific accuracy levels with a relatively short, user-tolerable calibration session.

Furthermore, an incremental data approach in a Bayesian framework offers a computational advantage over training linear models on pooled data. In addition, the shrinkage LDA model parameter updates were implemented in a much shorter amount of time with informative generic priors relative to updates with only user-specific data. Model recalibration mitigates the negative effects of data distribution drift during extended BCI use. A short model recalibration time could allow for more frequent continuous recalibration of BCI decoding models automatically with semi-supervised learning, e.g., [9], in the background without interruption during active BCI use.

Future work includes implementing sequential updates with linear models that enforce feature sparsity, such as stepwise LDA, as well as exploring fine-tuning of generic deep learning models, such as EEGNet [10].

REFERENCES

- [1] Farwell LA, Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*. 1988;70(6):510–523.
- [2] Gupta A, Liu A, Haines E, Alghamdi R, Collins LM, Mainsah BO. A data-centric analysis of the impact of training data quality vs. quantity on p300 brain-computer interface performance. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Jan. 2026.
- [3] Lin R et al. Assessing the impact of population data domain differences on transfer learning in p300-based brain-computer interfaces. In: *Proceedings of the AAAI*

Conference on Artificial Intelligence. Apr. 2025, 29415–29417.

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

[5] Townsend G et al. A novel p300-based brain-computer interface stimulus presentation paradigm: Moving beyond rows and columns. *Clinical Neurophysiology*. 2010;121(7):1109–1120.

[6] Krusienski D, Sellers E, McFarland D, Vaughan T, Wolpaw J. Toward enhanced p300 speller performance. *Journal of Neuroscience Methods*. 2008.

[7] Blankertz B, Lemm S, Treder M, Haufe S, Müller KR. Single-trial analysis and classification of erp components — a tutorial. *NeuroImage*. 2011;56(1):814–825.

[8] Murphy KP. Machine learning: A probabilistic perspective. MIT Press: Cambridge, MA (2012).

[9] Chen XJ, Collins LM, Mainsah BO. Language model-guided classifier adaptation for brain-computer interfaces for communication. In: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2022, 1642–1647.

[10] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*. 2018;15(5):056013.