Evolving Training Sets for Improved Transfer Learning in Brain Computer Interfaces

Jason Adair, Alexander Brownlee, Fabio Daolio, and Gabriela Ochoa

Computing Science and Mathematics
University of Stirling, Stirling, Scotland UK
{jad, sbr, fda, goc}@cs.stir.ac.uk

Abstract. A new proof-of-concept method for optimising the performance of Brain Computer Interfaces (BCI) while minimising the quantity of required training data is introduced. This is achieved by using an evolutionary approach to rearrange the distribution of training instances, prior to the construction of an Ensemble Learning Generic Information (ELGI) model. The training data from a population was optimised to emphasise generality of the models derived from it, prior to a re-combination with participant-specific data via the ELGI approach, and training of classifiers. Evidence is given to support the adoption of this approach in the more difficult BCI conditions: smaller training sets, and those suffering from temporal drift. This paper serves as a case study to lay the groundwork for further exploration of this approach.

Keywords: optimisation, machine learning, ensemble, brain-computer interface, p300, evolutionary computation, transfer learning

1 Introduction

Brain Computer Interfaces (BCI) are applications in which neurological recordings are utilised for the control of digital systems. Uses for BCI range from manipulation of prosthetic limbs, psychological interventions, and assisted communication devices [1]. Approaches to obtain these recordings can be separated into two main groupings; invasive and non-invasive. While invasive recordings can allow exceptional spatial and temporal resolutions, they involve sub-cranial surgery with potentially severe health risks and prohibitive financial costs [2]. For these reasons, non-invasive approaches have garnered significant interest. These include *electroencephalography (EEG)*; a technique that involves placing electrodes on the surface of the scalp to measure electrical fields produced by the underlying neurons. While this technique comes with little or no health risks, it lacks high resolutions [3], and is subject to noise from muscle movements (*electromyography*), cardiac rhythms (*electrocardiography*), eye movements (*electrooculography*), and environmental electrical sources [4].

Due to the aforementioned issues, a large quantity of training data is often required for each individual to calibrate the classifiers. As training sessions are often supervised by a health or technical expert, this increased training time not only comes at a financial cost, but proves frustrating and stressful for the participant which, in turn, introduces further noise into the training set. For these reasons, it is deemed imperative to minimise the amount of training data by exploiting all available data, from all potential sources.

We propose a novel method for the optimisation of the distribution of instances within a database of sets recorded from previous participants, in a manner that ensures that they can be used to create an ensemble that is maximally general to the population. This database is then used to seed a previously established method (Ensemble Learned for Generic Information) that recombines instances obtained from different participants with small quantities of participant-specific data, to create a robust participant-specific ensemble. This should allow for the creation of a BCI that requires only a small amount of training data, and should retain accuracy over time in a way that a traditional BCI does not. This is achieved by moving instances between previously obtained datasets via a random mutation hill-climber.

The structure of the paper is as follows: A brief literature introduction to Transfer Learning in the BCI field is given (Section 2), and algorithms described, with a hypothesis based on the new technique (Section 2.2). This leads us to the paradigm, dataset, and methodology used for experimentation (Section 3). Finally, the results are presented (Section 5 and 4) and discussed (Section 6).

2 Related work on Transfer Learning In BCI

As described in Section 1, BCIs are difficult to calibrate due to recordings having a low signal to noise ratio. This is further compounded by the non-stationary nature of brain signals: neural patterns not only differ between participants, but are also subject to *temporal drift*, where data obtained from a single participant changes drastically over time [5]. *Zero Training systems*, trained exclusively on participants from previous sessions, are an ideal goal, but this non-stationarity means highly accurate zero training systems may not be possible. Consequently, we must instead focus on minimising the participant-specific training information required by maximising the effectiveness of the data available.

Sufficient data from an individual for the creation of an accurate system comes with significant costs, so utilising databases from other participants offers an attractive avenue to alleviate this burden. Transfer Learning has been employed in a number of domains containing multiple sources to allow data inference to unseen sources. For a more in-depth discussion of the wider field, [6] provides a recent, thorough survey. More specifically, BCI literature typically reports *domain adaptation* approaches [5], the most popular of which being *Common Spatial Patterns* [7]. This involves creating a transformation of the data that will allow a single classification rule to be applied across all instances. A much less commonly explored approach is *'Rule Adaptation'* [5], in which a number of rules are created from the existing datasets, and then applied to the new instances. Both cases however, rely upon the natural distribution of the data as grouped by their original participant. Some attempts have been made to group datasets by known variants such as gender [8], and others using the information extracted from the trained models [9]; but little has been done in regards to instance selection for each model.

2.1 Ensembles

One method for incorporating data from other domains is the use of *ensembles*. Ensembles typically consist of an array of different classifiers trained with the same dataset. Each classifier makes predictions on a test set, and these are collated in a voting process. This allows multiple different relationships to be detected for the classification process, many of which may not be obvious, even to a domain expert. Another approach is to use multiple instances of the same classifier, trained with different initial datasets.

Ensembles have been used in a number of different BCI applications to increase accuracy and reduce the amount of training data required for participants. Arguably, the most well known *P300-Speller ensemble* is [10] in which an ensemble of SVMs were used to reduce variability in signal inputs by averaging classifier outputs, but relied on a substantial quantity of subject-specific data. This, like most BCI ensembles [11], used naive partitioning in which the instances were divided by their associated labels, whether it be by source domain or by stimuli. This proves useful for weighting classifiers within the ensembles; allowing information regarding the appropriateness of each model and the test-domain to be extracted [9]. It was demonstrated in [11] that overlapping these naive divisions can actually increase accuracy, suggesting that having the same training data duplicated amongst the classifiers can benefit the overall performance.

2.2 ELGI

In 2015, Xu et al [12] introduced the *Ensemble Learning Generic Information (ELGI)* approach. Rather than using the small amount of training data to train a classifier, or for weighting the models within a larger ensemble trained on the data of other participants, ELGI combines the participant-dependent data with participant-independent data to form a hybrid ensemble. This is achieved by splitting the datasets of each existing patient within the database into target and non-target sets. The removed missing instance class (target or non-target) is then replaced by a copy of the corresponding class from the participant-specific training data. This results in an ensemble consisting of $2n - 1$ classifiers, where n is the number of participants within the database.

This paper proposes a new technique in which the database containing the previously recorded participants' datasets are optimised to create an ensemble that is maximally generalised for the population, prior to the combination process of ELGI. The procedure is outlined fully in Section 4.

3 Methodology

This section defines the BCI Paradigm used in the work and describes the datasets. It then goes on to describe the offline filtering applied to the data and finally defines the algorithms to be compared in the experiments. In particular, we focus on how the datasets are initially derived for the eELGI approach, but background information on the application is summarised here for convenience and the interested reader can find more detail in [12].

3.1 P300 Speller Paradigm

A promising application of BCI systems are their use in communication assistive technologies. Some conditions, such as *Amyotrophic Lateral Sclerosis (ALS)*, cause degradation of physical movement [2], rendering patients unable to communicate with the outside world. Detection of neural activity can allow patients to control computers, and produce synthetic speech [1]. A common form of the system involves using a computer screen displaying a 6x6 grid of alphanumeric characters. The user concentrates on the character they wish to select, and all columns and rows are flashed in a randomised sequence. When the target character's row and columns flash, a fluctuation in neural patterns can be observed. This is known as a *P300 wave*. The goal of a BCI system in this context is to identify the P300 wave among the many other detected signals.

3.2 Dataset Recordings

The dataset used in this paper was obtained from [13], in which the P300 Speller Paradigm was adapted to present 6 images to the participant; each eliciting a response by increasing the brightness of the image. It included EEG recordings for 4 disabled participants and 4 able-bodied PhD students. Participants 1, 2 and 4 were able to speak with some dysarthria, but participant 3 was unable to communicate verbally due to their late stages of amyotrophic lateral sclerosis. All 4 disabled participants were wheelchair users, with limited to no control over their upper limbs. Each participant attended 4 recording sessions, each consisting of, on average, 810 trials; resulting in approximately 3240 trials per participant.

3.3 Prefiltering

The data underwent prefiltering as described in [13]. In summary, the procedure for this was: referencing against the mastoid electrodes, Butterworth bandpass filtering between 1 and 12 Hz, downsampling to 32 Hz, and Windsorizing to mitigate EoG and EMG artifacts.

3.4 Classifier

A Bayesian Linear Discriminate Analysis classifier (as in [13]) was used. Each stimuli presentation was treated as a binary problem, and the Bayesian probability of the prediction was recorded. Due to the paradigm structure, every subdivision of 6 stimuli presentations has 1 target and 5 non-target. These groupings are deemed as a ‘round’. A prediction is made based on the highest probability within each round. In each run, 20 rounds of all 6 stimuli are presented. This allows the Bayesian probabilities of each round to be summed with previous predictions, increasing predictive accuracy over the course of the run.

3.5 Conditions

The complex nature of BCI allows a number of different factors to be considered:

Quantity of Participant-Specific Data As a primary aim in BCI is minimising the required participant-specific training data, the impact of training set size was explored. The datasets follow a common hierarchical structure; each participant recording 4 sessions of 6 runs. All models were trained with data from the first session and 3 training set sizes were used: 3, 4, and 5 runs.

Time Between Testing Sessions A major challenge in BCI, other than between-participant transference, is between-session transference in single participants. As neural drift occurs over time, highly fitted models tend to lose accuracy. All models were tested on data acquired from 3 sessions, recorded over 2 days; session 2 on the same day as the training data, and 3 and 4 on a day no more than 2 weeks later.

3.6 Compared Algorithms

Three approaches were compared in our experiments, two taken from the literature and the proposed new method:

Standard Learning Individual Information (SLII) A Bayesian LDA model trained using participant-specific data exclusively. The highest probability in each round was selected as the target, and the rest, assumed to be non-targets[12].

Ensemble Learning Generic Information (ELGI) The ELGI [14] creates an array of classifiers by utilising the participant-specific and participant-independent datasets in the following manner:

$$[C_{2N}] = \sum_{i=1}^{N} [C(P_i^T + P_k^{NT}), C(P_i^{NT} + P_k^T)]$$

The training data P from each participant P_i is split into two subgroups; target T and non-target NT. A copy of the target instances from the test-participant

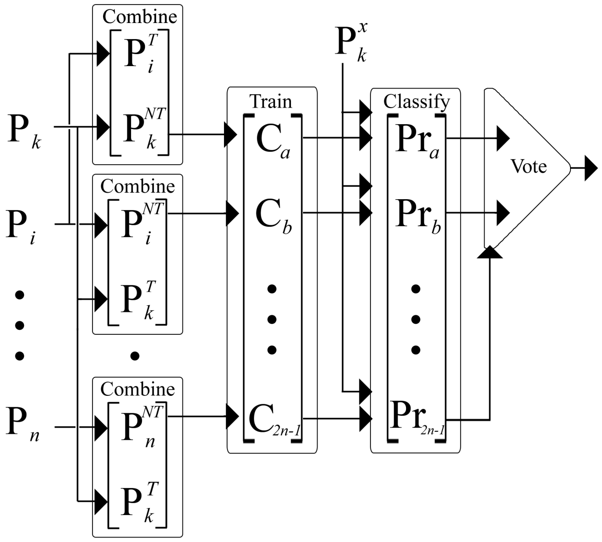


Fig. 1. ELGI approach displaying that 2 classifiers are trained for every participant in the database P_i by a splitting and recombination of their target P_i^T and non-target P_i^{NT} instances with the corresponding instances from the test-participant's training data P_k. These classifiers are then used to make predictions on the test-participants unseen data P_k^x. Finally, these predictions are collated via voting.

k (P_k^T) are then added to the non-target subgroup P_i^{NT}, and conversely, a copy of the test-participant's non-target instances P_k^{NT} are added to the target subgroup P_i^T. Each of these new subgroups are used to train an ensemble of classifiers C. Predictions Pr are made by each classifier in the ensemble based on the unseen data from the test-participant P_k^x, and these predictions are collated. This is done using the Sum Rule voting method where the Bayesian posterior probabilities are summed for each class. This is further depicted in Fig. 1.

Evolved Ensemble Learning Generic Information (eELGI) The novel proposed approach of this paper, as described in Section 4. In this, we assume that the natural grouping of instances by participant is not optimal. Instead, an evolutionary algorithm transplants instances between datasets taken from each participant, aiming to maximise the generalisability of each set in reference to other previously recorded participants, prior to their combination with participant-specific data via the ELGI.

4 Evolved ELGI Ensemble

We propose a new approach whereby the database containing the previous participants' datasets is optimised, with the goal of creating an ELGI ensemble that better generalises to the population. This is achieved by a leave-one-out technique in which a participant's bin, that is the subset containing all data from that participant, is selected at random, and a portion of the instances obtained from that participant are moved into the bin of another randomly selected participant. Two models are then trained; one using the data from the bin that was selected for transfer, and one from the bin that was selected as the destination. These models make predictions on the data in the remaining unselected bins. The resulting overall predictive accuracy is used as the fitness function for a random mutation hill climber. This seeks the allocation of training data to bins that maximises the predictive accuracy within the database.

We now describe the implementation in more detail. The procedure is given formally in Algorithm 1. The search is seeded with a solution consisting of 7 bins; each consisting of an individual's data, but excluding any information from the new participant, as in the Zero Training Model. A 500 iteration Hillclimber was then applied with the following mutation operator and fitness function.

Mutation (Move Operator) The move operator selects a target bin a and a destination bin b at random from the training set bins; a subset m with 10% of the target bin's instances are moved into the destination bin. Subsets P_{ea} and P_{eb} are created by removing subset m from P_a and appending it to P_b, respectively.

Fitness Function To assess the fitness of the candidate solution, 2 classifiers C_{ea} and C_{eb} were trained from the subsets P_{ea} and P_{eb}. These were then used to make predictions on the remaining instances within all subsets P, excluding the participant datasets selected for mutation (P_a and P_b). The average round accuracy over all the non-selected bins was calculated for both models affected by the mutation (f_{ea} and f_{eb}); a solution was deemed successful if the fitnesses obtained were an increase over the fitness (f_a and f_b) of both models created from the incumbent solution (C_a and C_b). The mutation was rejected if it caused a decrease in accuracy within either model.

This evolved dataset was then used to seed the original ELGI from [12].

5 Results

Figure 2 presents the performance of the SLII, ELGI and eELGI algorithms averaged across all 8 participants. Rows 1, 2 and 3 show performance of models with 3, 4 and 5 runs (see Section 3.5) of training data available, respectively. Columns display performance over 3 different testing sessions. While the confidence intervals of the different approaches vary due to differing sample sizes, the SLII and ELGI are almost indiscernible. The mean line of the eELGI is typically higher than that of the other algorithms, with its smaller confidence interval often visibly higher. The instances in which notable improvements are made are in the extremity conditions: low availability of participant specific data (row 1) and the testing session farthest from the training session (column 3).

Algorithm 1 Evolution of instances in eELGI

```
Input: Initial solution is $P = \mathcal{P}(P_i)$
Output: Final solution is Modified $P = \mathcal{P}(P_i)'$
for $x = 1 \rightarrow 500$ do
    Choose $a$ and $b$ from $1 : N$ where $N$ is the $|P|$
    Create $m \subset P_a$
    $P_{ea} \leftarrow P_a$ with $m$ removed
    $P_{eb} \leftarrow P_b$ appended with $m$
    Train classifiers $C_a$ and $C_b$ with $P_a$ and $P_b$
    Train classifiers $C_{ea}$ and $C_{eb}$ with $P_{ea}$ and $P_{eb}$
    $f_a = 0$, $f_b = 0$, $f_{ea} = 0$, $f_{eb} = 0$
    for $i = 1 \rightarrow N$ do
        if $i \neq a$ && $i \neq b$ then
            $f_a = f_a + C_a(P_i)$, $f_b = f_b + C_b(P_i)$
            $f_{ea} = f_{ea} + C_{ea}(P_i)$, $f_{eb} = f_{eb} + C_{eb}(P_i)$
        end if
    end for
    if $f_a < f_{ea}$ && $f_b < f_{eb}$ then
        $P_a = P_{ea}, P_b = P_{eb}$
    end if
end for
```

The Round Accuracy is presented in Figure 3 for the SLII, ELGI and eELGI algorithms. It is displayed by participant with each point representing the accuracy achieved with 3, 4 and 5 runs of training data provided for training. Increases in the quantity of participant-specific training data increases the predictive accuracy in each participant except 6. Participant 5 is the outlier in terms of variance; increases in participant-specific training data makes a much more substantial change to this classifier's accuracy than others. When considering overall round accuracies across differing training set sizes, eELGI performed better than the SLII and ELGI in 62.5% of cases, and obtained the second best results in the remainder. In no cases was eELGI the worst performer.

Figure 4 demonstrates each algorithm's resilience to neural drift over time. The round accuracy of the SLII, ELGI and eELGI over each of the testing sessions is given. A decrease in predictive accuracy was observed between session 2 and session 3 in 62.5% of the cases, and a decrease between session 3 and 4 in 58.3%. Overall, a decrease in predictive accuracy between session 2 and 4 was observed in 79.2% of the cases, as expected due to temporal neural drift. For 5 of the 8 participants, the eELGI retained the highest round accuracy after 2 weeks, while still maintaining relativity high accuracy in the remaining 3.

To analyse the differences between each algorithm's effectiveness in mitigating the effects of neural drift over time, hierarchical linear models were used as recommended in [15]. The results of these are given in Figures 5a and 5b. In Fig. 5a, lines show the expected average behaviour when considering the variation across participants, with points representing the residual deviation of each participant from the estimated common behaviour. Although no statistical sig-

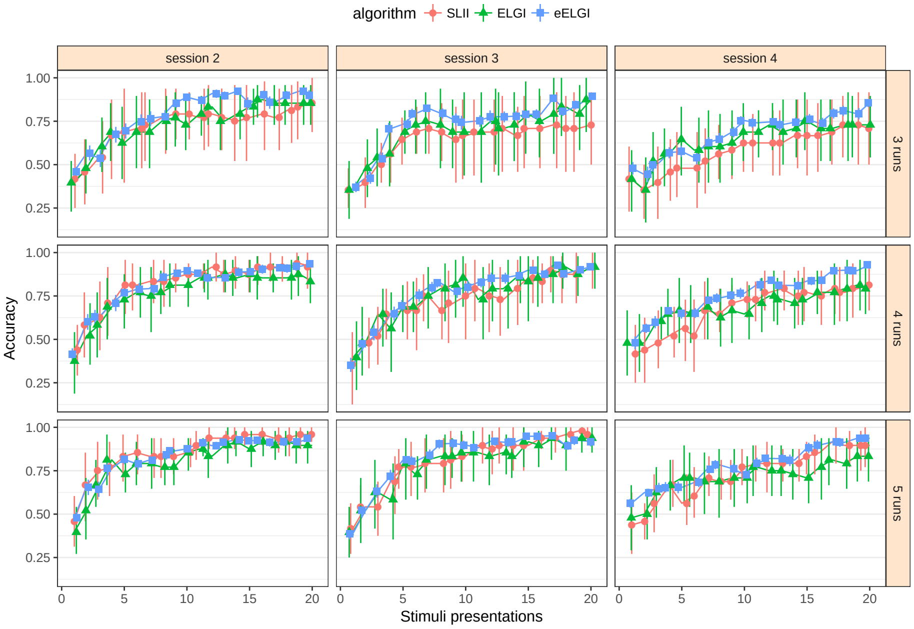


Fig. 2. Algorithm performance by number of stimuli presentations, with differing quantities of participant-specific training data available. Error bars show the confidence intervals around the means. Horizontal jitter has been added to improve discernibility.

nificance can be claimed here, the trends suggest that in all 3 testing sessions, the eELGI performed better than both the SLII and ELGI. It should also be noted that there appears to be less variance within and between testing sets for the eELGI. This suggests that the eELGI not only performs better than the other algorithms, but is also less susceptible to neural drift over time.

As seen in Fig. 5b, the round accuracy of all 3 algorithms increases with the amount of participant-specific data available. The SLII is most dependent on the quantity of participant-specific data, with ELGI performing much better when fewer training instances are available. However, this advantage is lost as volume of training data increases. The eELGI line has a similar slope to the ELGI (0.0402 and 0.0394, respectively) but with a higher intercept (0.618 to 0.574), resulting in better overall performance than both the SLII and ELGI in all 3 conditions. In fact, a post-hoc Tukey's comparison of the model estimates, averaging over algorithm-data interactions [16], showed that the eELGI produced a statistically significant increase in round accuracy over the SLII ($p = 0.0387$) while the ELGI did not ($p = 0.1483$). Therefore, with respect to the ELGI, the effect of evolving the base dataset appears to increase the intercept, without having any adverse affects to the rate of improvement seen when increasing participant-specific data.

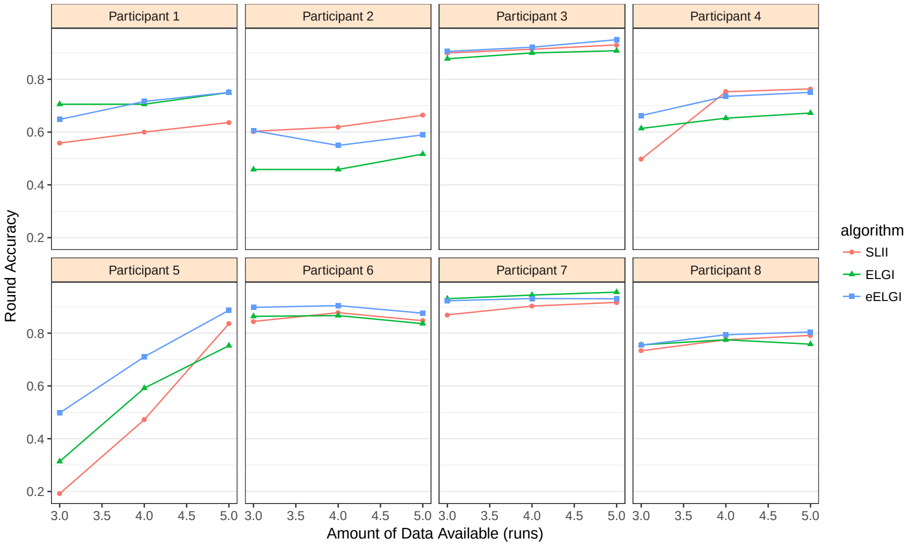


Fig. 3. Round Accuracy over all testing sets displayed for each quantity of participant-specific training data, separated for each participant.

6 Discussion and Conclusion

This paper served as a case study for the proposed eELGI approach. However, statistical significance can be difficult to determine with small datasets. This being said, even with small samples, we have demonstrated that there is a visible advantage to optimisation of the participant database for use in transfer learning techniques. We can see that an evolved database has 3 primary advantages:

1. *A higher classification accuracy*, regardless of quantity of training data. As seen in Fig. 3, 62.5% of cases see eELGI performing better than ELGI and SLII, with the remaining still close to the optimal. In Fig. 5b we can notice, in the majority of cases, a marked improvement over the non-evolved ELGI.

2. *A reduction in variance* in performance across not only sessions, but participants as well. When comparing sessions in Fig. 5a, and training set size in Fig. 5b, the groupings of round accuracies are noticeably more dense. Fig. 2, is perhaps the most dramatic demonstration of this. By including all participants over all test sets, the error bars for both the SLII and ELGI are substantial, while the eELGI provides a modest difference.

3. *A means for protection against temporal drift.* Fig. 5b demonstrates that the traditional BCI approach (SLII) is highly susceptible to the neural drift seen over time. While ELGI alleviates that to a degree, eELGI provides a much more linear, and slower degradation in predictive accuracy over the testing sessions.

As this paper focused on a small dataset, with an equal number of able and disabled patients, further work should investigate the effects of optimising different base datasets. For example, it should contain substantially more partic-

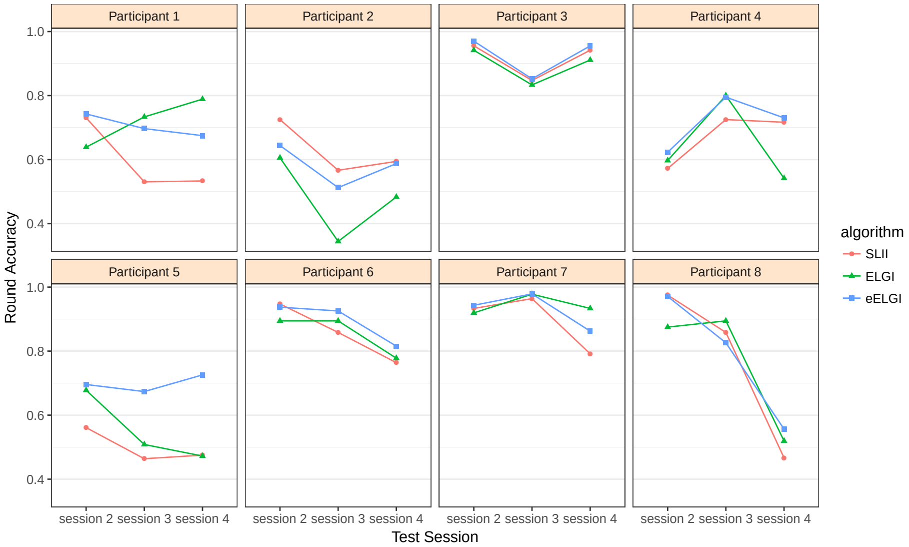


Fig. 4. Round Accuracy over all quantities of training data for each testing set, separated for each participant.

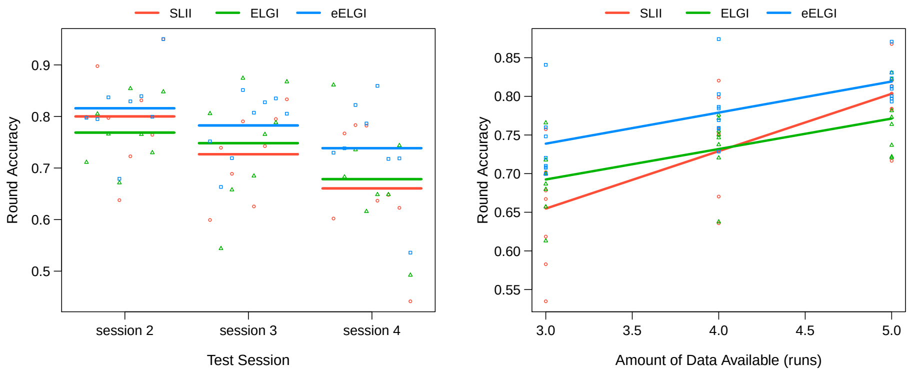
ipants, and, in more commonly observed situations, contain disproportionately more able bodied participants. In terms of algorithms; while a simple hillclimber has provided some promising results, it would be prudent to apply more complex heuristics to the problem. A potentially promising direction would be utilisation of a genetic algorithm with an encoding that would allow oversampling of the more prototypical instances.

Acknowledgements This research was funded by the ESPRC through the DAASE project [grant number EP/J017515/1]. The authors are grateful for the assistance of Kate Howie in preparating the statistical analyses.

Data Access Statement The dataset and source code used in this paper are available on request from the lead author.

References

1. Waytowich, N.R., Lawhern, V.J., Bohannon, A.W., Ball, K.R., Lance, B.J.: Spectral transfer learning using information geometry for a user-independent brain-computer interface. Frontiers in Neuroscience **10**(SEP) (2016)
2. Nicolas-Alonso, L.F., Gomez-Gil, J.: Brain computer interfaces, a review. Sensors **12** (2012) 1211–1279
3. Schwartz, A.B., Cui, X.T., Weber, D.J., Moran, D.W.: Brain-controlled interfaces: movement restoration with neural prosthetics. Neuron **52**(1) (oct 2006) 205–20



(a) Accuracy over time from training. (b) Accuracy over available training data.

Fig. 5. Fit of hierarchical linear models, with random effects for each participant, estimating (a) the overall Round Accuracy per testing set and (b) the change in Round Accuracy over training set size.

4. Khatwani, P., Tiwari, A.: A survey on different noise removal techniques of EEG signals. **2**(2) (2013) 1091–1095
5. Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., Grosse-Wentrup, M.: Transfer Learning in Brain-Computer Interfaces. IEEE Comp Intell Mag **11**(1) (2016) 20–31
6. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Volume 3. Springer International Publishing (2016)
7. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.R.: Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Processing Magazine **25**(1) (2008) 41–56
8. Cantillo-Negrete, J., Gutierrez-Martinez, J., Carino-Escobar, R., Carrillo-Mora, P., Elias-Vinas, D.: An approach to improve the performance of subject-independent BCIs-based motor imagery allocating subjects by gender. Biomed eng **13**(1) (2014)
9. Lotte, B.F.: to Minimize or Suppress Calibration Time in Oscillatory Activity-Based Brain Computer Interfaces. Proceedings of the IEEE (2015)
10. Rakotomamonjy, A., Guigue, V.: BCI Competition III: Dataset II- Ensemble of SVMs for BCI P300 Speller. IEEE Trans Biomed Eng **55**(3) (mar 2008) 1147–1154
11. Onishi, A., Natsume, K.: Overlapped partitioning for ensemble classifiers of P300-based brain-computer interfaces. PLoS ONE **9**(4) (2014)
12. Xu, M., Liu, J., Chen, L., Qi, H., He, F., Zhou, P., Cheng, X., Wan, B., Ming, D.: Inter-subject information contributes to the ERP classification in the P300 speller. Int'l IEEE/EMBS Conf. on Neural Engineering **2015-July** (2015) 206–209
13. Hoffmann, U., Vesin, J., Ebrahimi, T., Diserens, K.: An efficient P300-based brain-computer interface for disabled subjects. J Neurosci Methods **167** (2008) 115–125
14. Xu, M., Liu, J., Chen, L., Qi, H., He, F., Zhou, P., Wan, B., Ming, D.: Incorporation of Inter-Subject Information to Improve the Accuracy of Subject-Specific P300 Classifiers. International Journal of Neural Systems **26**(3) (2016) 1–12
15. Locascio, J.J., Atri, A.: An overview of longitudinal data analysis methods for neurological research. Dement Geriatr Cogn Dis Extra **1**(1) (2011) 330–57
16. Hothorn, T., Bretz, F., Westfall, P.: Simultaneous inference in general parametric models. Biometrical Journal **50**(3) (2008) 346–363