Predicting memory from study-related brain activity

Sucheta Chakravarty[1*], Yvonne Y. Chen[2] and Jeremy B. Caplan[1,3]

[1]Department of Psychology, University of Alberta, Edmonton, AB T6G 2E9. [2]Department of Neurosurgery, Baylor College of Medicine, Houston, TX 77030. [3]Neuroscience and Mental Health Institute, University of Alberta, Edmonton, AB T6G 2E1.

Author Note

Corresponding author: Sucheta Chakravarty, sucheta@ualberta.ca, Department of Psychology, Biological Sciences Building, University of Alberta, Edmonton, AB, T6G 2E9, Canada, Tel: +1.780.492.5361.

The authors thank Russ Greiner for helpful suggestions early in the project.

Supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Abstract

To isolate brain activity that may reflect effective cognitive processes during the study phase of a memory task, cognitive neuroscientists commonly contrast brain activity during study of later-remembered versus later-forgotten items. This "subsequent memory effect" method has been described as identifying brain activity "predictive" of memory outcome. However, the modern field of machine learning distinguishes between descriptive analysis, subject to overfitting, and true prediction, that can classify untrained data. First, we tested whether classic event-related potential signals were, in fact, predictive of later old/new recognition memory (*N*=62, 225 items/participant); this produced significant, but small predictive success. Next, pattern classification of the multivariate spatio-temporal features of the single-trial EEG waveform also succeeded in predicting memory. However, the prediction was still small in magnitude. In addition, topographic maps suggested individual differences in sources of predictive activity. These findings suggest that on average, brain activity, measured by EEG, during the study phase is only marginally "predictive" of subsequent memory. It is possible that this predictive approach will succeed better when other experimental factors known to influence memory outcome are also integrated into the models.

Keywords: subsequent memory effect; recognition memory; prediction; pattern classification; event-related potential

Predicting memory from study-related brain activity

Introduction

To analyze brain activity underlying successful memory formation, cognitive neuroscientists have adopted the so-called “subsequent memory effect” (SME; Sanquist, Rohrbaugh, Syndulko, & Lindsley, 1980), contrasting brain activity during the study phase of a task for subsequently remembered (hits) versus forgotten (misses) items. The SME is a major advance over prior methods that compared activity between different encoding conditions rather than relating it to eventual memory outcomes (for reviews of the SME, see Wagner, Koutstaal, & Schacter, 1999; Paller & Wagner, 2002; H. Kim, 2011). The SME approach has produced several highly replicated findings, including the late positive component (LPC) and the slow wave (SW) of the event related potential (ERP) of the EEG, both more positive for subsequent hits than misses (e.g., Sanquist et al., 1980; Karis, Fabiani, & Donchin, 1984; Chen, Lithgow, Hemmerich, & Caplan, 2014; Fabiani, Karis, & Donchin, 1990; A. S. Kim, Vallesi, Picton, & Tulving, 2009; Friedman, 1990; Smith, 1993). The robustness of the SME could be due to the fact that it indexes brain activity which is coupled with behaviour. For example, the Levels of Processing concept (Craik & Lockhart, 1972) holds that the likelihood of an item being remembered depends on the conceptual depth with which the participant evaluates or interacts with the item during encoding. Evidence has suggested different SME ERPs reflect these different processing-levels (Sanquist et al., 1980; Paller, Kutas, & Mayes, 1987; Fabiani et al., 1990).

Notably, the SME is often described as identifying brain activity “predictive” of memory success (Brewer, Zhao, Desmond, Glover, & Gabrieli, 1998; Wagner et al., 1998). If truly predictive, the SME could form the basis of important learning applications, such as tracking learning progress or testing the effectiveness of different training protocols (Fukuda & Woodman, 2015; Arora et al., 2018). Here, we examine whether “predictive” is an accurate characterization of SME ERP signals. Consider that in the traditional approach, SME ERPs are analyzed by the difference in brain activity at study for subsequent hits and misses, averaged across many trials. The hits–misses contrast is tested for statistical significance, with participants as repeated-measures. This captures the association between study phase brain activity and memory outcome, but such a descriptive model is not aimed at explaining the causal relationship or making predictions about new observations. To evaluate whether or not memory outcomes can, indeed, be predicted from the SME ERPs, it is important to apply predictive models, which remain under-explored in this context. Predictive models could be particularly helpful in bridging the gap between the existing theories and the potential learning applications.

Some recent studies suggest that prediction of memory from study activity could succeed with fMRI (Watanabe et al., 2011; Lee, Brodersen, & Rudebeck, 2013) as well as with intracranial EEG recordings (Weidemann et al., 2019; Weidemann & Kahana, 2019; Ezzyat et al., 2017; Arora et al., 2018). With standard, scalp-recorded EEG, Fukuda and Woodman (2015) showed that amplitude of the frontal slow wave and occipital alpha-band power, two pre-identified SME EEG measures, could predict the old/new confidence ratings given by the participants at test. Although this is a valuable finding, it skips predicting the memory outcome itself, our current aim. To classify subsequent memory from *multivariate* EEG activity at study, Noh, Herzmann, Curran, and de Sa (2014) used two classifiers. One was trained with pre-stimulus spectral features. The other used during-stimulus features, and was further divided into a time-domain and a spectral-domain classifier. Overall classification accuracy was near 60%, but the authors did not report the success rate, alone, of the time-domain signal during-stimulus. On the other hand, using only time-domain features (pre- and post-stimulus onset), Sun et al. (2016) found success in predicting subsequent memory for a majority of their participants ($N = 9$) with a convolutional neural network classifier [mean accuracy: 72.07%], whereas linear classifiers failed to classify the majority. Unlike linear classifiers, non-linear classifiers (such as convolutional neural networks) can evaluate the interactions between features, which could have led to this difference in success rates. Notably, the number of examples available to build or train these models in this type of context is usually very small for such interactions to be captured reliably. Also, it is possible that in some cases, the authors may have tried various variations of the classifier and for various reasons (including length limitations), only reported the best outcomes. This could result in inflated apparent success rates of classifiers (Skocik, Collins, Callahan-Flintoft, Bowman, & Wyble, 2016). In sum, for scalp-recorded EEG, it remains unknown if highly replicated ERP SME features are predictive . This is an important step in estimating a “benchmark” of predictive strength for this purpose. It is also unclear if the multivariate time-domain EEG signal can be used predictively. One possible limitation of the time-domain features is sensitivity to trial-to-trial variability in latency (Luck, 2014), which could impact classifier training. Thus, it is possible that researchers have tried and failed in the time domain, and opted to stop pursuing this goal in favour of spectral features (i.e., a file-drawer problem).

In the present study, we seek to understand the general prospect of using study related EEG time-domain features to predict memory with the help of easy to interpret, linear predictive models. First, using concepts from signal-detection theory (Green, Swets, et al., 1966), we ask if it is possible to predict memory outcome (i.e., hit or miss) for each study item based on individual, previously identified SME ERPs (mean amplitude of the LPC or the SW). We consider the probability distribution of the SME ERP for all the hits versus misses and test for the amount of separability between these two distributions which can support predictions for individual trials. The predictions are made based on the rule that hits are more positive than misses (for LPC or SW, as per prior findings) and by varying the classification threshold across the two distributions.

Importantly, the SME ERPs were identified through trial-averaging and planned comparisons, which could limit their use to predict memory. Trial-averages in the traditional, descriptive analysis can help raise the signal-to-noise ratio (SNR). But while this step can identify portions of the signal with low variability across trials, it can also wash out components with greater variability, which could also carry meaningful information related to memory-encoding. With planned comparisons, the electrodes of interest are based on prior studies, thus, possibly missing out on other relevant sources of activity. To move beyond these limitations, our next step was to use multivariate measures that include features beyond those known from the traditional research. The multivariate features were then analyzed with predictive models borrowed from the machine-learning literature. These models can automatically learn useful patterns from multivariate measures (Norman, Polyn, Detre, & Haxby, 2006) and are trained and tested on separate sets of data to evaluate its generalizability; a practice that checks for over-estimation of a model and is not, in general, looked at in descriptive analyses. Thus, with this approach, we can ask the more general question: does brain activity during the study phase predict memory at the test phase?

Another motivation for comparing the univariate predictors with multivariate, machine-learning classifiers was that the standard old/new recognition task is, arguably, impoverished. These kind of judgements are highly likely to be driven by more than one process, which are reflected in more than one neural activity measure. The multivariate classifiers are designed precisely for problems such as this (multiple predictors). These have the potential to discover multiple processes, and produce a combined prediction based upon this process-impure signal. Note that it is also possible to request additional subjective judgements, such as remember/know distinctions or confidence levels, to isolate multiple processes that are thought to underlie memory retrieval, such as recollection versus familiarity. This has been frequently done in previous classifier approaches to study phase as well as test-phase activity (Noh et al., 2014; Fukuda & Woodman, 2015; Noh, Liao, Mollison, Curran, & de Sa, 2018; Liao, Mollison, Curran, & de Sa, 2018; Sun et al., 2016). However, there is the risk that this approach may alter the way participants approach the task (Eldridge, Sarfatti, & Knowlton, 2002; Hicks & Marsh, 1999). Also, this relies on subjects' ability to cleanly separate their own familiarity versus recollection processes (see Dunn, 2008, whose the state-trace analysis casts a doubt on such ability).

Moreover, the issue of process impurity likely goes far beyond the recollection/familiarity distinction. Thus, to avoid this, our task included simple instructions for the participants (simply to study for a later memory test) and a simple response (old versus new). This also welcomed subject variability that could reveal interesting individual differences.

Regarding the recollection/familiarity distinction, specifically, the common dual-process view of ERPs in recognition-memory paradigms is that the FN400, an ERP elicited during the test phase of the task, reflects familiarity-based retrieval and the Late Parietal Positivity (LPP), also elicited during the test phase, reflects recollection-based retrieval (e.g., Rugg & Curran, 2007). In our data set, both of these signals produced significant old/new effects (see Chen et al., 2014), suggesting that both familiarity and recollection processes appeared at test. This confirms the process-impurity of the task. However, the FN400 (contrasting hits versus misses) covaried significantly with performance (d' and negatively with response time) across participants, whereas the LPP did not. This suggests that the putative recollection process, although clearly present, played a far more minor role in driving the old/new judgement than the putative familiarity (or conceptual priming; Voss & Paller, 2009) process. This led to clear predictions for the current classifier approach. Because the LPC SME covaried significantly across participants (Chen et al., 2014), with both the FN400 and performance (d' and response time), we expected that the LPC would produce above-chance classification of subsequent hits versus misses. Because the SW SME covaried significantly across participants with the LPP, but not with performance measures, we expected that the SW would not be able to classify subsequent old/new recognition above chance. Alternatively, variability reflected by the SW might be unrelated to individual differences, but could still support classification above chance when attempted within-subjects, as we do here.

Materials and Methods

Behavioural materials and procedure

Data were from the 64 participants for whom the traditional analysis of ERPs was previously reported in Chen et al. (2014). Of these, two participants were excluded for having more than 15% of the total number of study trials (225) rejected due to artifacts. The research was approved by a University of Alberta ethical review board.

The experiment involved alternating study and test phases (Figure 1). Participants were given a very simple instruction— to study the words for later tests. No instructions related to study strategies were provided. For each study list, they were instructed that they will see 25 words that are to be studied. Participants were not required to make any response during study. For each test list, they were asked to make old/new judgments by pressing the relevant key (1 for old, 2 for new). Words were presented one at a time, both at study and at test. Each study word was displayed on the screen for 1500 ms with a jittered inter trial interval (300–500 ms). Each study list consisted of 25 words and was followed by a short math distractor task, consisting of 5 addition or subtraction problems involving integers from 1 to 9. The math problem remained on the screen until the participant made a response. Each test list immediately followed the math distractor task and consisted of 50 words, 25 of which were from the study (i.e., “old”) and 25 were lures or “new” words. Each test word remained on the screen until the participant pressed either key. Hits were correctly responded study trials and misses were incorrectly responded study trials.

EEG methods

EEG was recorded in an electrically shielded, sound-attenuated chamber, from high-density 256-channel Geodesic Sensor nets (Electrical Geodesics Inc., Eugene, OR). Signal was amplified at a gain of 1000 and was sampled at 250 Hz (impedance below 50 $k\Omega$ and referenced to the vertex electrode, Cz). EEG signal was pre-processed with the EEGLAB toolbox (`http://sccn.ucsd.edu/eeglab`; Delorme & Makeig, 2004), running in

MATLAB. It was bandpass filtered to 0.5–30 Hz and average re-referenced. Independent component analysis (ICA) was used to look for artifacts in the signal (such as eye blinks, muscle noise etc.). EEG trials were then epoched from 100 ms pre-stimulus to 1200 ms post-stimulus intervals. After removing the baseline, we used a voltage threshold of 50 μV to remove epochs with large drifts. Additionally, for each epoch, we calculated the difference in voltage between adjacent time samples or the point-to-point difference to detect artifacts. We rejected epochs for which the point-to-point difference exceeded 25 μV. With the voltage and point-to-point difference thresholds in place, more than 15% of epochs were rejected for two participants (16% and 43% epochs rejected, respectively). Data from these two participants were excluded. For all other participants included in this study ($N = 62$), on average, 2 out of 225 epochs were rejected [min = 0, max = 23].

EEG Classification

We seek a function f that can predict discrete class labels Y (hit or miss in this case) to each trial X, i.e., $f(X) = Y$. X is a $N \times T$ matrix where N denotes the number of electrodes and T denotes the number of voltage samples as a function of time; $N \in \mathbb{Z}$ and $T \in \mathbb{R}$. Elements of X are called "features." Thus, f transforms the high dimensional space of EEG features onto a one-dimensional decision space. f is called a classifier. First, we tested if classic SME ERPs, such as the LPC or the SW, when computed for individual trials, are able to predict memory outcome for individual trials better than chance. Next, we tested if multivariate pattern analysis of EEG trials from the study phase can predict memory outcomes.

Classification based on SME ERPs. Two study-related ERPs were considered, consistent with prior research dating back to Karis et al. (1984): the LPC and the SW, from the centro-parietal electrode Pz. LPC is positive going, occurs between 400–700 ms after stimulus onset and more positive for hits than misses. SW is relatively sustained activity, occurring between 700–1200 ms. Across different SME studies, SW is reported for both centro-parietal and frontal electrodes. But frontal SW is thought to reflect item–item associations (A. S. Kim et al., 2009) or processing of emotional stimuli (Diedrich, Naumann, Maier, Becker, & Bartussek, 1997; Simon-Thomas, Role, & Knight, 2005). Because we used isolated common nouns, we did not expect to see it. The SW was subdivided into an early (700–900 ms post-stimulus) and a late (900–1,200 ms) component (see Chen et al., 2014).

For each SME ERP and for each study trial, we calculated the mean amplitude from electrode Pz, over the respective time window . The classification rule or function, based on prior (descriptive) SME results, was that subsequent-hits should have more positive voltage than subsequent-misses (Chen et al., 2014; Karis et al., 1984). Then, the receiver operating characteristic (ROC) curve was traced by setting each observed mean amplitude value as a classification threshold and plotting true positives (subsequent hits that were greater than or equal to the threshold) against false positives (subsequent misses that were greater than or equal to the threshold). After obtaining the ROC, the area under the curve (AUC) of the ROC was calculated through trapezoidal numerical integration implemented by the perfcurve function in MATLAB R2018a (see Figure 2 for a demonstration). AUC indexed the capability of the classifier to make more hits and less false alarms. AUC $= 0.5$ would reflect random predictions (chance), and a perfect classifier would achieve AUC$= 1$. Also, in this case, AUC< 0.5 would indicate that subsequent misses were on average more positive than subsequent hits.

Multivariate classification. Here, we used multiple EEG features per trial, with the speculation that other study related EEG features (beyond the known SME measures) could also be informative for making memory predictions. Each EEG epoch had 257 electrodes, sampled at 250 Hz for 1200 ms, thus there were over 80,000 features per trial. For computational simplicity, we selected a subset of these. First, as correlations are very common across neighbouring electrodes, we selected a set of 10 electrodes that span the recording coverage (Figure 3). Second, we averaged the signal for each electrode over

100 ms time bins, from 0–1200 ms post-stimulus onset. The resulting EEG signal consisted of 10 spatial × 12 temporal = 120 features.

When using multiple EEG features to make predictions, the classification rule is not known a-priori (unlike as above). However, we can learn this through predictive modelling. We used two models, linear discriminant analysis (LDA; Fisher, 1936) and linear support vector machine (SVM; Cortes & Vapnik, 1995). In general, linear models are advantageous because these are easy to interpret; the weight of a feature in the model indicates its relative importance in the classification. Each model has a set of parameters, values of which are set through examples, also known as "training". Once trained, the model can generate examples on its own, thus it can be used for predictions for unseen examples, also known as "testing".

It is crucial to test the model on unseen examples; for the model could be too specific to the training examples, often by capturing the noise in it (also known as "overfitting") and thus cannot generalize. To reduce overfitting, the weights of the features in the model can be scaled, also known as "regularization." We used a regularized LDA classifier (fitcdiscr, MATLAB 2018a) where the covariance matrix was calculated as: $\hat{\sum}_\gamma = (1-\gamma)\hat{\sum} + \gamma \mathrm{diag}\hat{\sum}$, where $\hat{\sum}$ is the empirical, pooled covariance matrix for the two classes and γ is a regularization parameter, lying between 0 to 1. SVM uses support vectors to draw hyperplanes that discriminate between the two classes. The support vectors are examples (here, EEG trials) that maximize the distance between the classes. We can reduce the chance of overfitting of an SVM model by setting a penalty (the box constraint parameter of fitcsvm; MATLAB R2018a) for mis-classifying examples that are on the class boundary. The default value for box constraint is 1 and in general smaller values allow for more regularization. In this study, for all LDA models, we set $\gamma = 1$, i.e., the maximum. For SVM, since a fixed maximum or minimum for the box constraint parameter does not exist, we chose a value that is reasonably smaller than the default value of 1, we set box constraint = 0.05. Importantly, the choice of regularization parameter values were independent of the test sets used to evaluate performance of the classifiers. Note that it is also possible to tune the regularization parameters for individual models. However, this did not alter the results substantially (see Figure 6).

We tested model performance through 10-fold cross-validation. Trials were randomly split into ten equal-sized folds, with nine folds being used to "train" the model and the remaining fold to "test" it. This was repeated ten times, ensuring that each trial was once used as a test trial. Cross-validation folds were stratified, such that the number of examples for the two classes were the same across all training folds. For each trial in a test fold, the trained classifier computed a "score," which readily translates into the posterior probability of that trial belonging to each class. Probability estimates across all trials for a test fold were then sorted and set as thresholds for calculating the corresponding true-positive and false positive rates, to trace the ROCs and compute the AUC. Average AUC across the 10 test-folds was used as the final estimate of the classifier performance.

Note that classifier success can also be evaluated by "accuracy", calculated as: $\frac{(TP+TN)}{(P+N)}$; TP = true positives, TN = true negatives, P = positives and N = negatives. However, in our data, the two classes, hits and misses, were imbalanced (described in detail in the next paragraph). Since accuracy does not take false alarms into consideration, for imbalanced sets, it is possible to achieve very high accuracy when the classifier has a bias to predict the over-represented class. Thus, we did not use accuracy as the measure for classifier performance.

Class imbalance. In our data, hits were more frequent than misses, making the training sets class-imbalanced. This could lead to the classifier getting biased towards learning more about and predicting more frequently the over-represented class. If this was true, re-balancing the classes either by undersampling the over-represented class or by oversampling the under-represented class can be helpful. Due to the small sample size of our data (≤ 225 in total per participant) we did not use undersampling. Instead, we used the Synthetic Minority Oversampling Technique (SMOTE; Chawla, Bowyer, Hall, &

Kegelmeyer, 2002; Arora et al., 2018) to create new examples from the existing under-represented class examples. To create a new example, the algorithm 1) randomly selects an existing example from the under-represented class, 2) randomly selects one example from its k-nearest neighbours (from the same class), 3) calculates the distance between the two chosen examples, 4) adds a random number between 0 and 1 to this distance and 5) adds the distance (with added random noise) to the first chosen example. The new example, created this way, lies in between the original example and its chosen neighbour. We set the number of nearest neighbours in the SMOTE algorithm to 4 but note that the first nearest neighbour is the example itself. Thus, the effective number of nearest neighbours considered for each example was 3. Synthetic minority samples were computed until the total number of examples in the two classes matched. Importantly, we only used SMOTE to balance the training sets. If SMOTE is used to balance the entire dataset, it is possible to end up with very similar trials in the training and testing sets, creating a double-dipping problem.

Cluster analysis of LDA weights. For LDA, we can assess the importance of a feature from its coefficient or weight in the model. To check if any pattern existed in the distributions of feature-weights across participants, we performed a cluster analysis in MATLAB (R2018a) using the k-means algorithm (kmeans function from the Statistics and Machine Learning toolbox; Martinez, Martinez, & Solka, 2017). For a specified number of clusters, n, the algorithm minimizes the within class variance or the sum of distance of each point in a cluster from the centroid of the cluster. We ran the cluster analysis separately for 2, 3, 4 and 5 clusters. To avoid local minima, each clustering solution was minimized over 100 replications. For each clustering solution, we calculated the following distance measure for each participant (using the function silhouette in MATLAB R2018a):

$$S_i = \frac{(y_i - x_i)}{\max(x_i, y_i)},$$

where x_i is the average of all distances from the i^{th} participant in one cluster to all other participants in the same cluster and y_i is the minimum of the average distances from the i^{th} participant to all other participants in all clusters other than its own cluster. This measure can range from -1 (indicating probable wrong assignment of a participant in a cluster) to 0 (participant can belong to either of the neighbouring clusters) and up to 1 (participant is distant from the neighbouring clusters). A set of 2 clusters was found to be the best possible solution, with the highest average value for this measure (0.11) across all participants [see Figure 10 for a visual representation of the distance measures across all participants, separately for the cases of 2, 3, 4 and 5 clusters]. To visualize the feature-weight pattern for each cluster, we used spline-interpolated topographic plots, created by the topoplot function of the EEGLAB toolbox (Delorme & Makeig, 2004). An inverse distance-weighting interpolation was used. This means that feature weight values for electrodes that were not used in the classification, were calculated from the weighted averages of the same for the electrodes used in the classification.

All analyses were done using in-built and custom written functions and scripts in MATLAB R2018a. Specific functions from the Statistics and Machine Learning Toolbox (Martinez et al., 2017) were also used. Although the classification problem was set up for each participant individually, to gauge overall success of the methods, one sample t-tests (against chance level, 0.5) were done. We also carried out Bayesian t tests using a MATLAB function by SamPenDu (2015). The Bayes factor is the ratio of Bayesian probabilities for the alternative and the null hypotheses; $BF_{10} = \frac{p(H_1)}{p(H_0)}$. By convention (Kass & Raftery, 1995), $BF_{10} > 10$ provides strong evidence for the alternate and $BF_{10} < 0.1$ provides strong evidence for the null. For $BF_{10} > 3$ and $BF_{10} < 0.3$ there is some evidence for the alternate or the null, respectively. Effect sizes of the classifiers were estimated from the 95% confidence intervals. To ensure that our results can be reproduced over multiple runs of the scripts, a pseudo-random number generator algorithm was specified in MATLAB R2018a (Mersenne twister, seed = 0).

Results

We start with the traditional ERP analysis of the subsequent memory effect. Figure 4 presents these ERPs at electrode Pz, averaged across all participants ($N = 62$), whereby hits appeared to be more positive than misses. Paired t-tests between the mean voltage for hits and misses for the LPC was significant; $t(61) = 2.89,\ p < 0.05$. The difference was also significant for the early SW; $t(61) = 3.04,\ p < 0.005$. Note that these effects were comparable to those reported by previous studies. For example in Paller et al. (1987), the reported F ratio for the LPC was 8.6, thus the corresponding t-statistic can be estimated to be 2.93 (i.e., square-root of the F ratio), which is very similar to the present study. However, the ERP effect was not significant for the late SW; $t(61) = 1.82,\ p = 0.07$. We also calculated the Bayes factor, BF_{10}, which showed some evidence for the subsequent memory effect for the LPC ($BF_{10} = 6$) and the early SW ($BF_{10} = 9$) but was inconclusive for the late SW ($BF_{10} = 0.7$).

Next, we tested if these known SME ERP measures could predict subsequent memory for individual trials. Since the ERP effect for the late SW was not significant, we did not include it in this analysis. Accordingly, from here on, we refer to the early SW simply as SW. We found that for both ERP measures (Figure 5), AUCs (across all participants) were significantly above chance (0.5), $t(61) = 3.31,\ p < 0.005,\ BF_{10} = 18$ for LPC and $t(61) = 3.35,\ p < 0.005,\ BF_{10} = 19$ for SW. However, in each case, the 95% confidence intervals for the AUCs were close to chance; [0.51 0.54] for both LPC and SW. Also, across participants, AUCs were significantly correlated between the LPC and SW measures, $r(60) = 0.65,\ p < 0.0001$, (Figure 7). This could be due to the general temporal auto-correlation property of the EEG signal. In sum, classification of single trials from the study phase using a-priori measures achieved small but significant success.

Next, we tested if multivariate brain activity from the study phase, as measured with EEG, could predict subsequent memory and if it can do so better than the individual SME ERPs. As noted in the methods section, we selected a set of 10 electrodes and 12 time-samples, i.e., 120 features in total. We used two linear classifiers: LDA and linear SVM, along with a stratified 10-fold cross validation technique. AUCs were averaged across the 10 folds. To reduce chances of overfitting, we used regularization; the regularization parameters were set to be constant across the models (also, optimizing these parameters for individual models did not alter our results, see Figure 6). Across participants, the AUCs for both LDA and SVM (Figure 5b, left) were significantly better than chance, $t(61) = 3.54,\ p < 0.001,\ BF_{10} = 33.54$ for LDA and $t(61) = 4.55,\ p < 0.0001,\ BF_{10} > 500$ for SVM. The corresponding 95% confidence intervals were [0.51 0.55] for LDA and [0.52 0.56] for SVM. Also, pairwise one tailed t-tests showed that SVM performance was significantly greater than the SME ERP based classifiers [SVM versus LPC: $t(61) = 1.83,\ p < 0.05, BF_{10} = 1.28$; SVM versus SW: $t(61) = 1.76,\ p < 0.05, BF_{10} = 1.13$]. However, this was not true for LDA [LDA versus LPC: $t(61) = 0.62,\ p = 0.27, BF_{10} = 0.24$; LDA versus early SW: $t(61) = 0.70,\ p = 0.24, BF_{10} = 0.26$]. Given that the multivariate models had more degrees of freedom than the SME ERP based classifiers, these results suggest that overall, the time domain EEG signal during the study phase is only marginally predictive of subsequent memory success. Moreover, predictive success was positively correlated between LDA and SVM (Figure 8a), $r(60) = 0.74,\ p < 0.0001$, suggesting that participants who were easier to classify by one method were also easier to classify by the other.

Notably, for both LDA and SVM, a small subset of participants were found to have AUCs far below chance (see Figure. 8a). This is possible, for the assumption of a symmetric null distribution for the classifier performance may not hold in the case of small sample size data with small effect size (Jamalabadi, Alizadeh, Schönauer, Leibold, & Gais, 2016). In that case, non-parametric tests may be better suited. Following up on this, for each participant we conducted a Mann Whitney U test between the AUC values for all of the 10 folds and chance (0.5). Then, we calculated the z transform of the U statistic. Finally, we used t-tests to check if the z scores across all participants were significantly differently from zero. This showed that the z-scores for both LDA and SVM were significantly positive, [LDA: $t(61) = 2.55,\ p < 0.05, BF_{10} = 3$, SVM: $t(61) = 3.13,\ p < 0.005, BF_{10} = 11$]. This confirms that even if the assumption of symmetry for the null distribution is relaxed, the LDA and SVM classifiers in our study were overall better than chance.

Imbalanced classes might have challenged classifier training. Alternatively, participants with better memory may have a greater signal-to-noise ratio (SNR) that the classifier could identify. Across participants, a weak positive trend (Figure. 9a–b) was observed between AUCs and the proportion of hits. This trend was significant for SVM, $r(60) = 0.38, p < 0.005$ but not for LDA, $r(60) = 0.21, p = 0.09$. We also calculated the sensitivity index or d' of participants performance, which showed similar results as the proportion of hits [LDA: $r(60) = 0.14,\ p = 0.29$; SVM: $r(60) = 0.28,\ p < 0.05$]. To investigate if the imbalance between the trial numbers for hits and misses influenced our classifier results, we balanced the trials by oversampling the misses with Synthetic Minority Oversampling Technique or SMOTE (see Methods; Chawla et al., 2002). AUCs (Figure 5b, right) were yet again significantly above chance [LDA: $t(61) = 3.13,\ p < 0.005,\ BF_{10} = 10.91$, SVM: $t(61) = 3.72,\ p < 0.001,\ BF_{10} = 57$]. However, the 95% confidence intervals were not better than that without oversampling (LDA: [0.51 0.54]; SVM: [0.52 0.55]). Predictive success remained positively correlated across LDA and SVM, $r(60) = 0.80,\ p < 0.0001$ (Figure 8b). Thus, while imbalanced classes often pose a challenge to classifier training, in this case, it could not account for the relatively small prediction rate. Instead, participants with better recognition memory appear easier to classify (see Discussion for implications of this). The positive trend between classifier performance and proportion of hits was also observed after the classifiers were trained with balanced classes [LDA: $r(60) = 0.14,\ p = 0.29$; SVM: $r(60) = 0.13,\ p = 0.31$]. For SVM, when participants with very low AUCs (< 0.45) were excluded, this trend was significant, $r(51) = 0.27,\ p < 0.05$. However, classifier performance did not correlate with d' in this case.

For participants with LDA AUCs above 0.5 ($N = 43$), we wondered which features were deemed more important by the classifier for the classification. A cluster analysis of the LDA feature-weights revealed two subgroups of participants with distinct patterns (see Methods and Figure 10). $N = 22$ participants were found to be in cluster 1 and $N = 21$ in cluster 2. Figure 11 shows the topographic plots for the LDA feature-weights, averaged across all participants in each cluster and for three different time windows: 0–100 ms, 501–600 ms and 1001–1100 ms (see Figures 16 and 16 for the full version, i.e., for all the time-windows). For cluster 1, for the very early 0–100 ms time window, greater feature weights were observed on the left and right parietal regions. On the other hand for cluster 2, for the same time window, greater feature weights were observed in the fronto-central region. Given that these earlier time windows are more likely to reflect perceptual processing, one possibility is that these differences in feature weight patterns are indicative of the potential difference in attentional mechanisms between the two clusters. For a later time window, 501–600 ms, which is closer to the onset of LPC activity, only cluster 1 showed greater weights for the central parietal scalp region, whereas for cluster 2, greater weights were observed more widely in the right frontal and parietal regions. Thus, the topographic plot for cluster 2, for the 501–600 ms did not resemble the posterior positivity feature observed in the SME ERP analysis of this data set (see Chen et al., 2014). For an even later time window, 1001–1100 ms, the patterns for the two clusters were almost orthogonal; cluster 1 showed greater weights in the left parietal region whereas cluster 2 showed greater weights in the frontal and slightly left parietal region. It is possible that this difference is indicative of potentially different spontaneous study strategies between the participants of the two clusters.

We were also curious as to whether the standard SME ERP effects might be different for the two clusters. Investigating this, follow up analysis of the corresponding ERPs at electrode Pz (Figure 12) showed a general trend for hits to be more positive than misses (i.e., the classic subsequent memory effect) for both clusters. However, this trend was clearly more pronounced for cluster 1 than cluster 2. We conducted a 2×2 ANOVA on mean LPC amplitude with the within subject factor memory success (hit versus miss) and between subject factor cluster (1 and 2). This revealed a significant interaction between the two factors, $F(1, 41) = 15.72,\ p < 0.001,\ \eta_p^2 = 0.28$, whereas the LPC effect was significant for cluster 1, it was not so for cluster 2. The same ANOVA design on the mean amplitude for SW, also showed similar results. The average LDA AUC for cluster 1 was similar to that of cluster 2 [mean ± SD of AUC for cluster 1 $= 0.56 \pm 0.05$; cluster 2 $= 0.57 \pm 0.04$] and the average proportion of hits was comparable between clusters 1 and 2 [mean ± SD for proportion of hits for cluster 1 $= 0.79 \pm 0.11$; cluster 2 $= 0.80 \pm 0.08$]. The d' values were also comparable between the two clusters [mean ± SD for d' for cluster 1 $= 2.13 \pm 0.62$; cluster 2 $= 2.09 \pm 0.87$]. Overall, this could suggest that there may be *at least* two different types of feature patterns that could form the basis for predicting memory.

Discussion

The subsequent memory approach is often referred to as identifying brain activity “predictive” of memory. However, limited attempts have been made to test this claim with actual predictive models. Here, using signal detection theory, we showed that two, previously identified, SME ERPs, namely, the LPC and the SW, could indeed predict memory (hit or miss) for individual trials in a word recognition task. However, across participants ($N = 62$), the success rate was small. Considering the SME approach is limited by many factors, such as planned comparisons and trial averaging, the small success may be expected. Also, multiple processes could be at play for memory judgments in a recognition task, each associated with different sources of neural activity. Thus, instead of single ERPs, analysis of patterns in the multivariate EEG waveform at study may fare better at predicting memory. To test this, we employed machine learning classifiers (LDA and linear SVM), which are well suited to analyze multivariate structures. These models were used to learn memory relevant patterns from a set of 120 spatio-temporal EEG features from individual study trials. Both LDA and SVM achieved significant success in predicting memory, albeit still with a small success rate. Since generalization was also accounted for in LDA and SVM with ten fold cross validation, the success of these models further strengthens the possibility of predicting subsequent memory from EEG activity at study. However, when comparing LDA and SVM performance with that of the LPC or SW based classification, only SVM showed a small significant improvement. Thus, despite the considerably greater degrees of freedom, these models did not offer an obvious improvement over classification with LPC or SW alone. But, interestingly, exploratory analysis on the features of importance to the LDA classifier showed that there were two subgroups of participants with seemingly different activity patterns. On average, one of these subgroups (cluster 1, see Figure 16) showed greater feature weights for the posterior scalp region, which is similar to the findings from the univariate ERP analysis of the same dataset (see Chen et al., 2014). It also agrees with previous SME ERP studies that have shown that memory success can be associated with a greater positive going signal over the parietal region (for a review, see Paller & Wagner, 2002),. However, for the other subgroup (cluster 2, see Figure 17), greater weights were observed in the frontal region. Further, post-hoc analysis of SME ERPs at electrode Pz, separately for the two subgroups, showed significant LPC as well as SW effects for cluster 1 but not for cluster 2. Interestingly, previous literature also suggests that a frontal slow wave may be invoked by associative processes whereas the posterior slow wave may reflect elaborate item-oriented processing (for e.g., see Kamp, Bader, & Mecklinger, 2017). Although we can not know this for sure, one possible reason for the involvement of the frontal region in cluster 2 could be that it reflects some associative strategies for learning, spontaneously undertaken by the participants in this group. Notably, it would not be possible to identify these subgroups without the classifier models. Thus, both the univariate and multivariate predictive analysis reported in the current study have their own merit. Below we discuss potential improvements and limitations towards predicting memory.

..

We sought out to understand the general level of challenge in predicting memory from EEG activity during the study phase. Unlike other approaches, where failed or less successful analyses might not be disclosed, so the degree to which best-cases are reported becomes impossible to judge, we report a systematic sequence of classification analyses, to avoid apparently inflated success rates. We did not exclude participants based on their performance in the task, which is commonly done (Noh et al., 2014; Sun et al., 2016; Watanabe et al., 2011). Thus, although the 95% confidence intervals of our classifier success are modest for the aggregate, the regression (Figure 9b) suggests meaningfully large classification success rates. Also, to avoid any possible circularity in the analysis, we did not select the multivariate features based on the univariate SME results (Coutanche, 2013; Noh et al., 2014). Instead, we selected these features based on the general EEG knowledge (scalp coverage, correlations etc.), which substantially minimizes the chance of over-estimating the effect. Thus, our results provide a benchmark for the effect size for this type of classification to be compared against. This could improve with more fine-grained analysis, for example through other feature selection or feature reduction methods or even with the help of non-linear classifiers including state-of-the-art neural networks. Including EEG spectrogram features, which are more resilient to trial-to-trial latency fluctuations, may also lead to better performance (Ezzyat et al., 2017; Weidemann et al., 2019).

Notably, class-imbalance (hits versus misses) was common in our data set and this could have biased the training of LDA and SVM towards the over-represented class (hits). However, re-balancing classes offered no improvement, allaying such concerns, at least in our case. Alternatively, it is possible that participants with more hits also have high-SNR brain activity, which could have helped the classifier. We found some support for the latter, as SVM performance increased significantly as the proportion of hits increased (for similar evidence, see Arora et al., 2018). Whenever memory was close to chance, the corresponding brain activity may have had less information for the classifier to pick up on.

Conversely, participants who performed better might, to some degree, have been those who (and whose brains) were more engaged in the task, producing higher task-relevance of their brain-activity.

Although we cannot know for sure why better-performing participants may be easier to classify, two causes come to mind. First, some lower-performing participants might have low motivation, a plausible possibility, given that participants did not self-select as research participants, but were recruited from via a course-based research participation pool, in exchange for partial course credit. There was no disincentive to speed through the experiment or disengage from the task. For such participants, brain activity may simply not be task-relevant. Second, participants who struggle with the task more, genuinely finding the task challenging, may have task-related brain activity that is more variable, or obscured by cognitive processes related to frustration, or strategic exploration, etc. Both causes could lead to lower SNR. In future studies, this could be addressed by pre-calibrating the task for each participant to equate difficulty across the sample, and increase the level of motivation across participants, for example, through rewards. Both these modifications might produce substantially higher levels of classifier success as well across the sample.

One may conclude that due to individual variability in performance and likely in brain activity too, a large sample size, as in our study, is essential to obtain overall significant results with the classifiers. To test this idea, we estimated the minimum sample size we might have needed for the classifier analysis to succeed. With bootstrap techniques, we tested significant effects for the SVM classifiers for different sample sizes, ranging from 6 to 62 participants, which were selected at random and without replacement. For each sample size, we generated 100 sets of participants and for each set we calculated one sample t-test to check if the corresponding SVM AUCs were significantly better than chance. Then, across the 100 sets, we calculated the average effect or the probability of obtaining AUCs that were not overall significantly better than chance. This showed that the probability to obtain a non-significant effect for SVM decreases very sharply with increasing sample size (Figure 13), up to about 30 participants. For sample sizes greater than 30, this probability is very close to zero. Thus, we were not after a result that is only made possible by a large sample size. In fact, in many cases, a sample size of about 15 participants may be enough to obtain significant results (probability for a non-significant effect < 0.5, see Figure 13), provided the number of trials per participant is high or at least comparable to our study. Thus, it is conceivable that Sun et al. (2016) failed to find overall success with simple linear classifiers due to small sample size ($N = 9$). Interestingly, while some of the early, influential, SME ERP studies do not pass this sample size ($N > 15$) criterion (Brewer et al., 1998; Karis et al., 1984; Neville, Kutas, Chesney, & Schmidt, 1986; Sanquist et al., 1980; Wagner et al., 1998), others do so (Smith, 1993; Otten & Rugg, 2001; Paller et al., 1987; Van Petten & Senkfor, 1996; Friedman, 1990). However, many of these also have considerably lower trial counts.

Many decades of behavioural research (Kahana, 2012; Humphreys, Bain, & Pike, 1989; Neath, 1998; Lewis, 1979) points to numerous factors that determine memory success, that should not be visible through the lens of study-related activity alone (for an alternate account, see Weidemann & Kahana, 2019). Examples include competition from other items at retrieval, nature of the retrieval task (such as recognition, free recall, serial recall, cued recall, word-stem completion, word-fragment completion, lexical decision), retrieval time, output encoding, rehearsal and response criterion (recognition tasks). The serial positions of the items in the studied list can also influence subsequent memory (for example, primacy and recency effects) and are possibly reflected in brain activity as well (Talmi, Grady, Goshen-Gottstein, & Moscovitch, 2005; Rushby, Barry, & Johnstone, 2002; Sederberg et al., 2006). However, these factors are usually not accounted for in the SME approach. In addition, the Encoding Specificity principle (Tulving & Thomson, 1973) suggests that remembering will be more successful when there is a good match of context between study and test than when they mismatch (for an alternate account, see Nairne, 2002). Context could be spatial/environmental, temporal or internal mental or physical state (Howard &

Kahana, 2002). This study–test contextual match is also overlooked with the SME. Brain activity indexed by the SME may also relate to experimental manipulations such as attention (Paller et al., 1987; Otten & Rugg, 2001; Summerfield & Mangels, 2006), intentional learning (Paller, 1990; Karis, Bashore, Fabiani, & Donchin, 1982), use of different learning strategies (Karis et al., 1984; Rugg & Curran, 2007), etc. Semantic congruity of the the to-be-remembered stimuli (Neville et al., 1986) as well as the type of the stimulus (for example, verbal, pictorial, abstract patterns etc., see Fabiani et al., 1990; Paller et al., 1987; Friedman, 1990; Van Petten & Senkfor, 1996) can also influence the SME. Also, study and test phases are temporally distinct, but some aspect of brain activity may co-vary across these two phases (Chen et al., 2014), and test activity (Rugg & Curran, 2007) is also an important determinant of memory. Brain activity at retrieval may even be more reflective of important determinants of memory success (Weidemann et al., 2019; Polyn, Natu, Cohen, & Norman, 2005), including item-distinctiveness (LaRocque et al., 2013). Clearly, memory encoding is multifaceted and thus, a more extensive model that incorporates different cognitive measures as well as measures of brain activity may be more effective in predicting memory (Halpern et al., 2018).

Accordingly, it is likely that our current classifiers are “under-performing” their potential. One important factor missing from the SME approach and possibly influencing our classifiers is that retrieval performance is, to a large extent, competitive. Thus, probability of remembering an item not only depends on the corresponding EEG activity for that item at study but also on the EEG activity during the study of other items. Also, over the course of an EEG recording session, there is usually drift in the signal, mainly due to the electrodes drying out or sliding. Additionally, it is also possible that as the task progresses, the participant takes up on or shifts their strategy or approach towards the task. All of these factors could influence the classification. To test this, with the LPC and SW classifiers, we calculated classification performance separately for each of the nine lists studied by each participant (Figure 14). Lists with all hits or all misses were not included.

Indeed, the classification improved as the task progressed. Linear regressions between average AUC (across participants) for each list and list number (1 to 9) were significant, for both LPC [$F(1,7) = 6.34,\ p < 0.05$] and SW [$F(1,7) = 7.35,\ p < 0.05$]. Similar trends may also be possible for the multivariate classifiers, but due to the very small number of trials available per list for training the models, we did not follow up on that. However, performance measures such as d' or the proportion of hits for individual lists did not vary significantly with list number.

Also, given the wealth of research on distinctions between recollection- and familiarity-based retrieval, one important future direction could be to incorporate those distinctions into the classification— as indeed, has been done by some previous studies (Noh et al., 2014; Fukuda & Woodman, 2015; Noh et al., 2018; Liao et al., 2018; Sun et al., 2016). As with incorporation of other relevant variables (previous two paragraphs), this could improve classification accuracy. Two classifiers could be trained, one to classify based on a familiarity-like signal and one based on a recollection-like signal. The two classifiers could then be combined to produce a higher overall classification success rate. However, the subjective responses distinguishing recollection versus familiarity might be variable, in themselves, and thus introduce noise into the classification. Moreover, Dunn (2008) showed that remember/know judgements, themselves, appear to be based upon a summation of recollection and familiarity evidence. Thus, alternatively, it could be more effective to let a multivariate classifier "discover" the two (or more) neural processes and their optimal summation weights.

Importantly, in the traditional ERP or other similar univariate analysis, brain activity is averaged over many trials to increase the signal to noise ratio (SNR). Then, measures from this averaged brain activity signal are computed for behavioural conditions of interest and are compared across participants. With this approach, we may be able to identify some components of brain activity relevant to that behaviour, it should also be considered that the brain itself does not compute such averages to produce the behaviour.

Instead, this is produced by the firing of networks of neurons. In that sense, the classifiers, which learn from the multivariate pattern of brain activity specific to individual events, may be closer to the way the brain works than the traditional approach. However, since the classifiers are driven by the data only, it is also possible for them to learn relations that are different from what actually produces behaviour. Thus, obtaining a function-to-structure map of memory may also be inaccessible with present methods of obtaining brain-activity measures (Henson, 2005). The two different clusters of participants identified here could reflect individual variability in studying the same information, for example, use of different strategies. Classifiers could also be showing differences due to the word stimuli (frequency, imageability etc.).

Also, if the EEG activity from study can predict memory, then, hypothetically, it could also be used to guide restudy, as attempted by Fukuda and Woodman (2015). But in their case, the restudy re-randomized the relationship between initial study-related EEG activity and eventual memory outcome. It is possible that when enforcing better learning for stimuli tagged as “likely-to-be-forgotten” by the classifier, stimuli that were initially more likely to be remembered become weakened. If the goal is only to be able to predict memory, it may be possible find differences that lead to some classifier success, which is appreciated. But, to be employed in a memory training framework, we may need to isolate EEG activity that taps into truly effective encoding processes.

In doing so it may also be possible to find memory relevant neural activity patterns that not only generalize within the trials for one participant but also across multiple participants. This is an interesting future direction and very few studies have attempted at “between subject” prediction of memory (Liao et al., 2018; Koch, Paulus, & Coutanche, 2020). Specific to EEG, Liao et al. (2018) found some success with the test phase activity. However, their experiment requested additional subjective judgements from the participants, such as remember–know as well as source and confidence judgements. Accordingly, the between subject classifiers were set up to predict memory outcomes that were constrained to these additional judgments rather than simple old/new responses. Thus, although their results can not be directly compared to the current study, we were curious if between subject prediction for hits versus misses is possible with the study phase EEG activity in our data set. However, given the small success rates of the within subject classifiers, we suspected that this may fail. We tested this with a leave-one-subject-out cross validation procedure. Data from one participant were selected at random and used as the test set. Data from all other participants were used to train the classifier. We chose linear SVMs, as our results show these may be better than the LDA. The leave-one-subject-out cross validation was repeated 62 times, i.e., until data from each participant were used as the test set. This produced AUCs greater than 0.5 for 34 out of 62 participants, i.e., for about 54% of the total sample, thus it was not significant across participants, $t(61) = 0.98,\ p = 0.33, BF_{10} = 0.22$.

Also, with fMRI, Koch et al. (2020) were able to predict the average encoding pattern, between their participants. Following this idea, we checked if we could predict the average EEG waveform at study for hit and miss events for a participant, based on the same for the other participants. Once again, we used leave-one-subject-out cross validation and linear SVM classifiers. Also, since in this case each test set has only two trials (averaged waveform for hit and miss), instead of calculating AUCs for individual participants, we pooled together the classifier scores across subjects to calculate the ROC and the AUC of the ROC. Further, we used 1000 bootstrap samples to calculate the 95% CI of the AUC. This produced an AUC of 0.64, along with a 95% CI of [0.54 0.74] (Figure 15). Thus, it was possible to predict the average waveform for hit and miss events between participants. However, this may not be too surprising as the variance of the miss waveform may, in general, be higher than the hit waveform, due to the disparity in their trial counts, as we have discussed before. The classifier may be able to learn based on this difference in variance. Regardless, this initial set of analyses suggest that there may be multiple interesting directions to follow up on in the future with between subject classifications.

In sum, SME ERPs such as the LPC and SW may not only be related to memory success at the aggregate level, but could also predict memory for individual trials, albeit with small effect size. Some increase in effect size was achieved by using more features of the study-trial activity, and through multivariate pattern classification. Methodological improvements to the classification analysis may be able to increase the performance even further (for example, by using more complex algorithms and/or spectrogram information) and will be addressed by future research. Also, it is possible that unlike the EEG signal, the SME measured by the fMRI may contain a better SNR to predict memory success for individual trials. Alternatively, it is also quite possible for the classification success to never approach the maximum possible outcome, due to the numerous cognitive factors that are known to significantly influence memory success, but are not directly taken into account in the subsequent memory approach. In that case, even a low, but above-chance, classification is important, and a small level of success is, in fact, expected.

References

Arora, A., Lin, J.-J., Gasperian, A., Maldjian, J., Stein, J., Kahana, M., & Lega, B. (2018). Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial eeg recordings. *Journal of Neural Engineering*, *15*(6), 066028.

Brewer, J. B., Zhao, Z., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1998). Making memories: brain activity that predicts how well visual experience will be remembered. *Science*, *281*(5380), 1185–1187.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, Y. Y., Lithgow, K., Hemmerich, J. A., & Caplan, J. B. (2014). Is what goes in what comes out? encoding and retrieval event-related potentials together determine memory outcome. *Experimental Brain Research*, *232*(10), 3175–3190.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Coutanche, M. N. (2013). Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us? *Cognitive, Affective, & Behavioral Neuroscience*, *13*(3), 667–673.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21.

Diedrich, O., Naumann, E., Maier, S., Becker, G., & Bartussek, D. (1997). A frontal positive slow wave in the erp associated with emotional slides. *Journal of Psychophysiology*, *11*, 71–84.

Dunn, J. C. (2008). The dimensionality of the remember–know task: a state-trace analysis. *Psychological Review*, *115*(2), 426-446.

Eldridge, L. L., Sarfatti, S., & Knowlton, B. J. (2002). The effect of testing procedure on remember–know judgments. *Psychonomic Bulletin & Review*, *9*(1), 139-145.

Ezzyat, Y., Kragel, J. E., Burke, J. F., Levy, D. F., Lyalenko, A., Wanda, P., … Kahana, M. J. (2017). Direct brain stimulation modulates encoding states and memory performance in humans. *Current Biology*, *27*(9), 1251–1258.

Fabiani, M., Karis, D., & Donchin, E. (1990). Effects of mnemonic strategy manipulation in a Von Restorff paradigm. *Electroencephalography and Clinical Neurophysiology*, *75*(1-2), 22–35.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Friedman, D. (1990). Cognitive event-related potential components during continuous recognition memory for pictures. *Psychophysiology*, *27*(2), 136–148.

Fukuda, K., & Woodman, G. F. (2015). Predicting and improving recognition memory using multiple electrophysiological signals in real time. *Psychological Science*, *26*(7), 1026–1037.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Halpern, D., Tubridy, S., Wang, H. Y., Gasser, C., Popp, P. O., Davachi, L., & Gureckis, T. M. (2018). Knowledge tracing using the brain. *International Educational Data Mining Society.*

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *The Quarterly Journal of Experimental Psychology Section A*, *58*(2), 193–233.

Hicks, J. L., & Marsh, R. L. (1999). Remember–know judgments can depend on how memory is tested. *Psychonomic Bulletin & Review*, *6*(1), 117-122.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: a theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*(2), 208.

Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C., & Gais, S. (2016). Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. *Human brain mapping*, *37*(5), 1842–1855.

Kahana, M. J. (2012). *Foundations of human memory.* Oxford University Press, USA.

Kamp, S.-M., Bader, R., & Mecklinger, A. (2017). Erp subsequent memory effects differ between inter-item and unitization encoding tasks. *Frontiers in Human Neuroscience*, *11*, 30.

Karis, D., Bashore, T., Fabiani, M., & Donchin, E. (1982). P300 and memory. *Psychophysiology*, *19*(3), 328–328.

Karis, D., Fabiani, M., & Donchin, E. (1984). "P300" and memory: Individual differences in the Von Restorff effect. *Cognitive Psychology*, *16*(2), 177–216.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kim, A. S., Vallesi, A., Picton, T. W., & Tulving, E. (2009). Cognitive association formation in episodic memory: Evidence from event-related potentials. *Neuropsychologia*, *47*(14), 3162–3173.

Kim, H. (2011). Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *NeuroImage*, *54*(3), 2446–2461.

Koch, G. E., Paulus, J. P., & Coutanche, M. N. (2020). Neural patterns are more similar across individuals during successful memory encoding than during failed memory encoding. *Cerebral Cortex*, *30*(7), 3872–3883.

LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spector, K., & Wagner,

A. D. (2013). Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. *Journal of Neuroscience*, *33*(13), 5466–5474.

Lee, A. C., Brodersen, K. H., & Rudebeck, S. R. (2013). Disentangling spatial perception and spatial memory in the hippocampus: a univariate and multivariate pattern analysis fmri study. *Journal of Cognitive Neuroscience*, *25*(4), 534–546.

Lewis, D. J. (1979). Psychobiology of active and inactive memory. *Psychological bulletin*, *86*(5), 1054.

Liao, K., Mollison, M. V., Curran, T., & de Sa, V. R. (2018). Single-trial eeg predicts memory retrieval using leave-one-subject-out classification. In *2018 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 2613–2620).

Luck, S. J. (2014). *An introduction to the event-related potential technique.* MIT press.

Martinez, W. L., Martinez, A. R., & Solka, J. (2017). *Exploratory data analysis with MATLAB.* Chapman and Hall/CRC.

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory*, *10*(5-6), 389–395.

Neath, I. (1998). *Human memory: An introduction to research, data, and theory.* Thomson Brooks/Cole Publishing Co.

Neville, H. J., Kutas, M., Chesney, G., & Schmidt, A. L. (1986). Event-related brain potentials during initial encoding and recognition memory of congruous and incongruous words. *Journal of Memory and Language*, *25*(1), 75–92.

Noh, E., Herzmann, G., Curran, T., & de Sa, V. R. (2014). Using single-trial EEG to predict and analyze subsequent memory. *NeuroImage*, *84*, 712–723.

Noh, E., Liao, K., Mollison, M. V., Curran, T., & de Sa, V. R. (2018). Single-trial eeg analysis predicts memory retrieval and reveals source-dependent differences. *Frontiers in human neuroscience*, *12*, 258.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

Otten, L. J., & Rugg, M. D. (2001). Electrophysiological correlates of memory encoding are task-dependent. *Cognitive Brain Research*, *12*(1), 11–18.

Paller, K. A. (1990). Recall and stem-completion priming have different electrophysiological correlates and are modified differentially by directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(6), 1021.

Paller, K. A., Kutas, M., & Mayes, A. R. (1987). Neural correlates of encoding in an incidental learning paradigm. *Electroencephalography & Clinical Neurophysiology*, *67*(4), 360–371.

Paller, K. A., & Wagner, A. D. (2002). Observing the transformation of experience into memory. *Trends In Cognitive Sciences*, *6*(2), 93–102.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*(5756), 1963–1966.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences*, *11*(6), 251–257.

Rushby, J. A., Barry, R. J., & Johnstone, S. J. (2002). Event-related potential correlates of serial-position effects during an elaborative memory test. *International Journal of Psychophysiology*, *46*, 13-27.

SamPenDu, D. F. (2015). Bayes factors Matlab functions.

Sanquist, T. F., Rohrbaugh, J. W., Syndulko, K., & Lindsley, D. B. (1980). Electrocortical signs of levels of processing: perceptual analysis and recognition memory. *Psychophysiology*, *17*(6), 568–576.

Sederberg, P. B., Gauthier, L. V., Terushkin, V., Miller, J. F., Barnathan, J. A., & Kahana, M. J. (2006). Oscillatory correlates of the primacy effect in episodic memory. *NeuroImage*, *32*, 1422-1431.

Simon-Thomas, E. R., Role, K. O., & Knight, R. T. (2005). Behavioral and electrophysiological evidence of a right hemisphere bias for the influence of negative emotion on higher cognition. *Journal of Cognitive Neuroscience*, *17*(3), 518–529.

Skocik, M., Collins, J., Callahan-Flintoft, C., Bowman, H., & Wyble, B. (2016). I tried a bunch of things: the dangers of unexpected overfitting in classification. *BioRxiv*, 078816.

Smith, M. E. (1993). Neurophysiological manifestations of recollective experience during recognition memory judgments. *Journal of Cognitive Neuroscience*, *5*(1), 1–13.

Summerfield, C., & Mangels, J. A. (2006). Dissociable neural mechanisms for encoding predictable and unpredictable events. *Journal of Cognitive Neuroscience*, *18*(7), 1120–1132.

Sun, X., Qian, C., Chen, Z., Wu, Z., Luo, B., & Pan, G. (2016). Remembered or forgotten?—An EEG-based computational prediction approach. *PloS one*, *11*(12), e0167497.

Talmi, D., Grady, C. L., Goshen-Gottstein, Y., & Moscovitch, M. (2005). Neuroimaging the serial position curve: a test of single-store versus dual-store models. *Psychological Science*, *16*(9), 716–723.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352.

Van Petten, C., & Senkfor, A. J. (1996). Memory for words and novel visual patterns: Repetition, recognition, and encoding effects in the event-related brain potential. *Psychophysiology*, *33*(5), 491–506.

Voss, J. L., & Paller, K. A. (2009). Remembering and knowing: electrophysiological distinctions at encoding but not retrieval. *NeuroImage*, *46*, 280-289.

Wagner, A. D., Koutstaal, W., & Schacter, D. L. (1999). When encodong yields remembering: insights from event-related neuroimaging. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *354*(1387), 1307–1324.

Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., . . . Buckner, R. L. (1998). Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, *281*(5380), 1188–1191.

Watanabe, T., Hirose, S., Wada, H., Katsura, M., Chikazoe, J., Jimura, K., … Konishi, S. (2011). Prediction of subsequent recognition performance using brain activity in the medial temporal lobe. *NeuroImage*, *54*(4), 3085–3092.

Weidemann, C. T., & Kahana, M. J. (2019). Neural measures of subsequent memory reflect endogenous variability in cognitive function. *BioRxiv*, 576173.

Weidemann, C. T., Kragel, J. E., Lega, B. C., Worrell, G. A., Sperling, M. R., Sharan, A. D., … Kahana, M. J. (2019). Neural activity reveals interactions between episodic and semantic memory systems during retrieval. *Journal of Experimental Psychology: General*, *148*(1), 1.

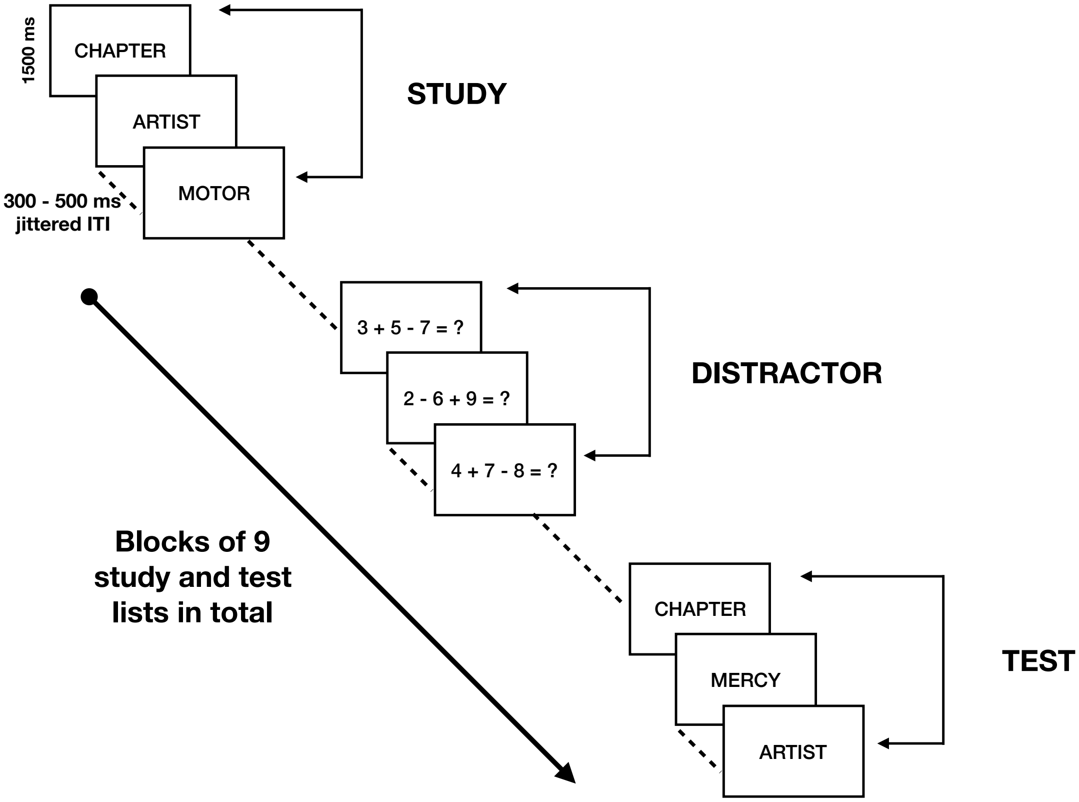

Figure 1. The experimental paradigm. Participants were asked to study a list of 25 words, presented one at a time at the center of the screen. This was followed by a short distractor task with simple math problems. Participants were then given a set of item recognition tests, judging each word as “old” (targets) or “new” (lures). There were equal number of targets and lures in the test phase. This whole process was repeated 9 times, yielding 225 study and 450 test trials. Each study list was unique. The order of the items during study was same as the order of the targets at test, with lures being presented at random positions in the list; lure items were not repeated across lists nor within lists.

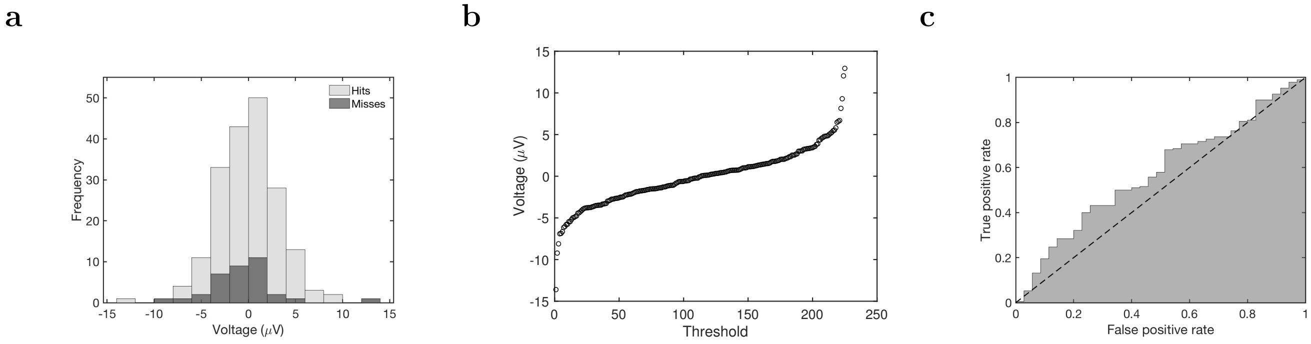

Figure 2. Demonstration of classification based on SME ERPs. a. Distribution of the LPC amplitude (from Pz) across all trials for a randomly selected participant. b. The thresholds used for classification. c. The ROC curve, shaded region represents the AUC. Dashed black line denotes chance.

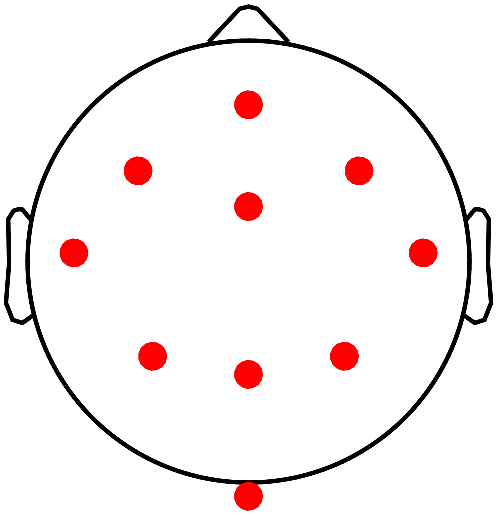

Figure 3. Selected electrodes for the multivariate classification, roughly distributed in equal between the frontal and posterior scalp regions.

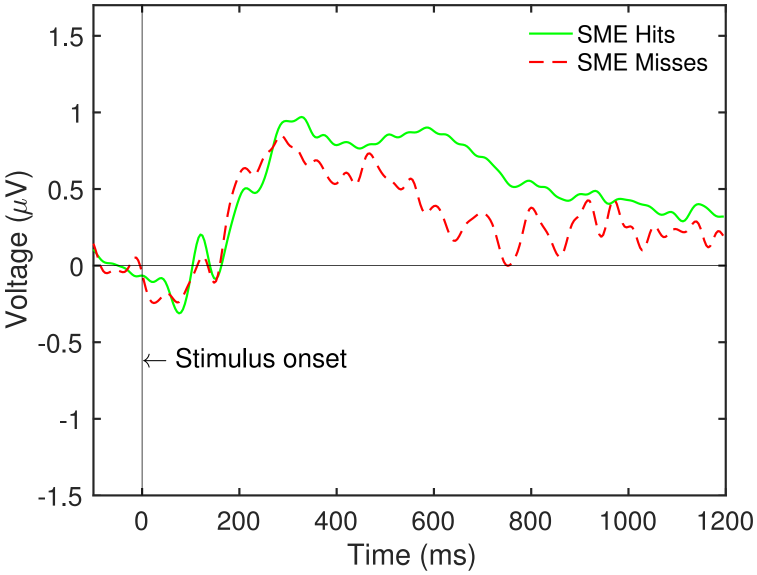

Figure 4. Grand averaged ERPs at electrode Pz for subsequently remembered (hits) and forgotten trials (misses).

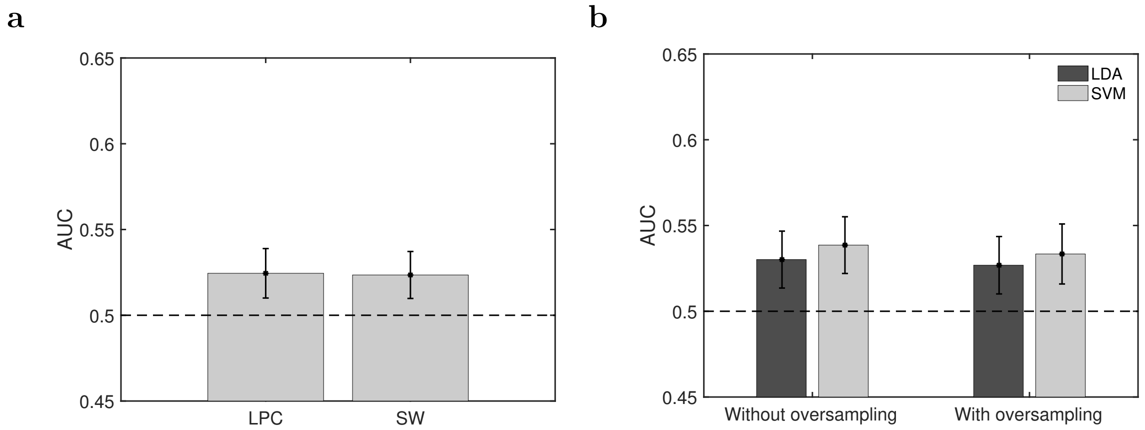

Figure 5. a. Classification based on SME ERPs: LPC and SW (computed from electrode Pz). Maximum AUC observed was 0.69 for both LPC and SW (for the same participant). b. Multivariate classification with LDA and SVM (left) and with oversampling to produce balanced classes (right). Maximum AUCs observed were 0.69 for LDA and 0.73 for SVM (same participant for LDA and SVM and also same as above). With balanced classes, maximum AUC for both LDA and SVM was 0.69 (same participant for LDA and SVM but different from above). Error bars are 95% confidence intervals. Dashed black line denotes chance level (0.5).

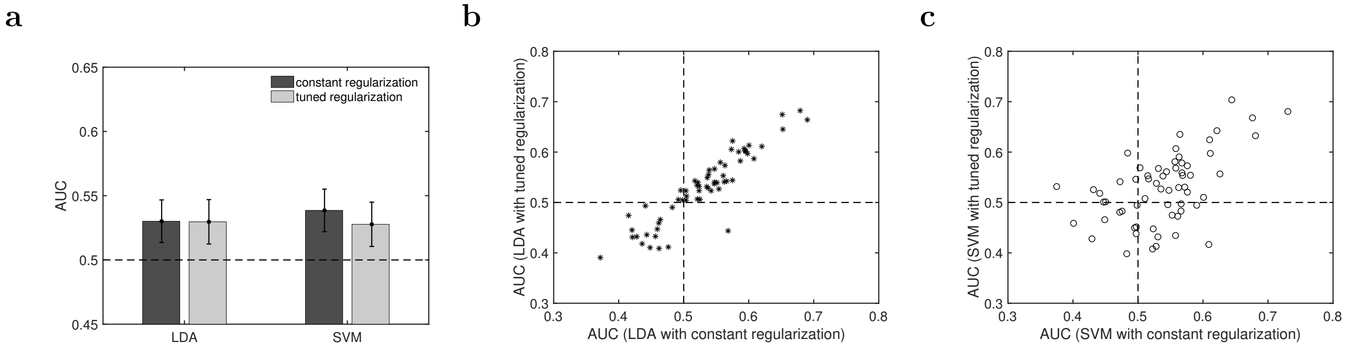

Figure 6. Effect of tuning the regularization parameters gamma of LDA and box constraint of SVM. We used a nested cross validation procedure. For the outer cross-validation, data was randomly partitioned into 10 stratified folds, 9 folds being used for training and 1 for validation. Then, the training data was subjected to an inner 9 fold stratified cross validation to tune the regularization parameter. For each training set of the inner cross validation, separate LDA models were trained for gamma = $[0,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9,\ 1]$. Similarly, for SVM, separate models were trained for box constraint = $[0.001,\ 0.005,\ 0.01,\ 0.05,\ 0.1,\ 0.5,\ 1,\ 5,\ 10,\ 50,\ 100]$. Then performance for these models were computed for the test folds of the inner cross validation. Value of the regularization parameter corresponding to the model with best performance was selected. Then this value was used in the model for the training data of the outer cross validation and then tested with the left out validation set. Finally, AUCs were averaged across the 10 validation sets. a. The overall effect of tuning the regularization parameters for each model. b. and c. AUCs for individual participants with constant and tuned regularization parameters.

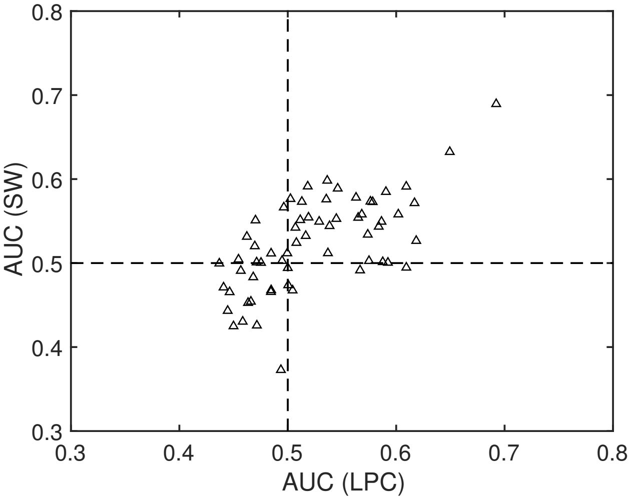

Figure 7. Correlation between AUCs for LPC and SW. Dashed lines denote chance.

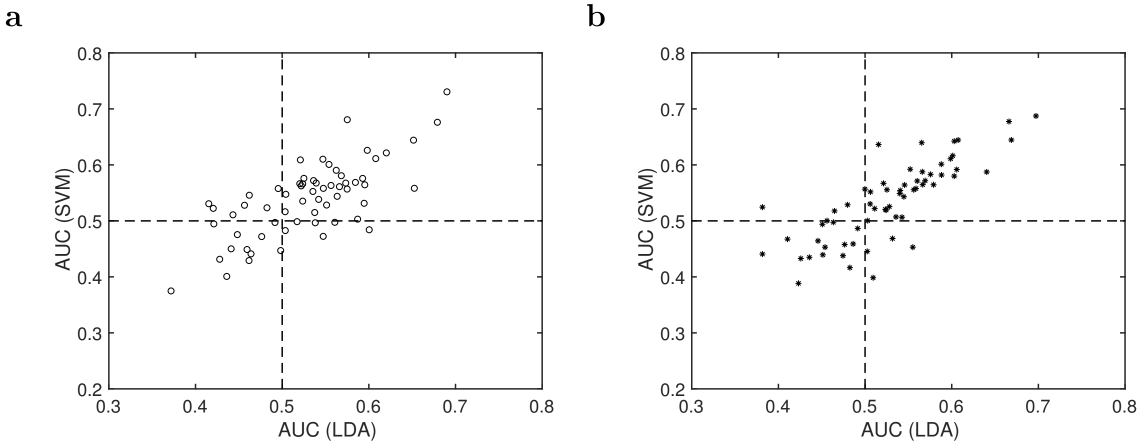

Figure 8. Correlation between AUCS for the two classifiers (LDA and SVM) with (a) and without (b) balanced classes for training. Dashed black lines denote chance.

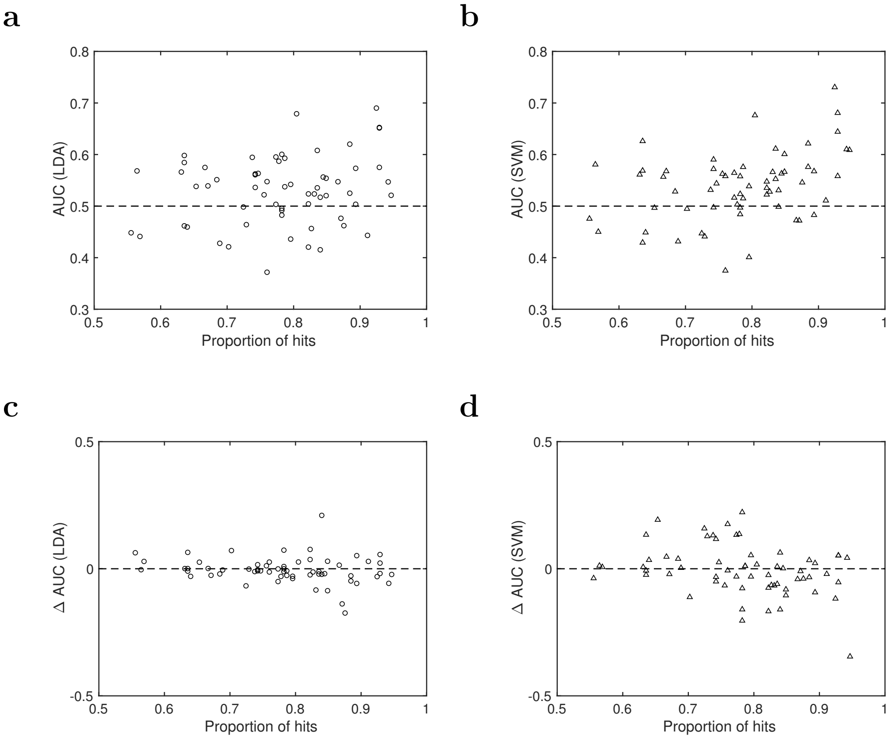

Figure 9. Relationship between classifier performance (AUC) and proportion of hits for LDA (a) and SVM (b). Percent change in classifier performance (ΔAUC) after oversampling, separately for LDA (c) and SVM (d).

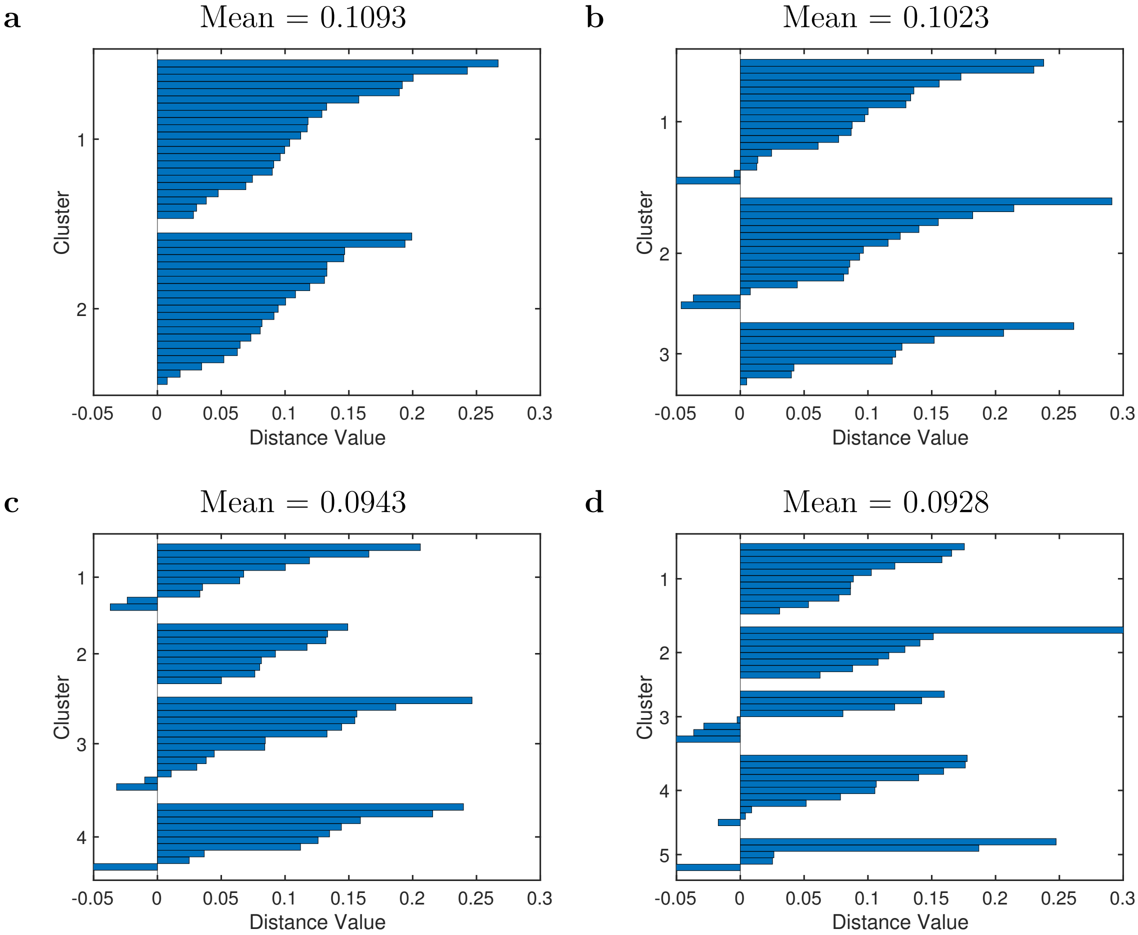

Figure 10. Determining the correct number of clusters for the cluster analysis of LDA feature weights. Each plot shows the distance measure for each participant for their respective clusters. Average distance scores across all participants are listed on top of the plot. For a set of two clusters (a), this measure was the highest. Also, all participants show positive distance scores for a set of two clusters.

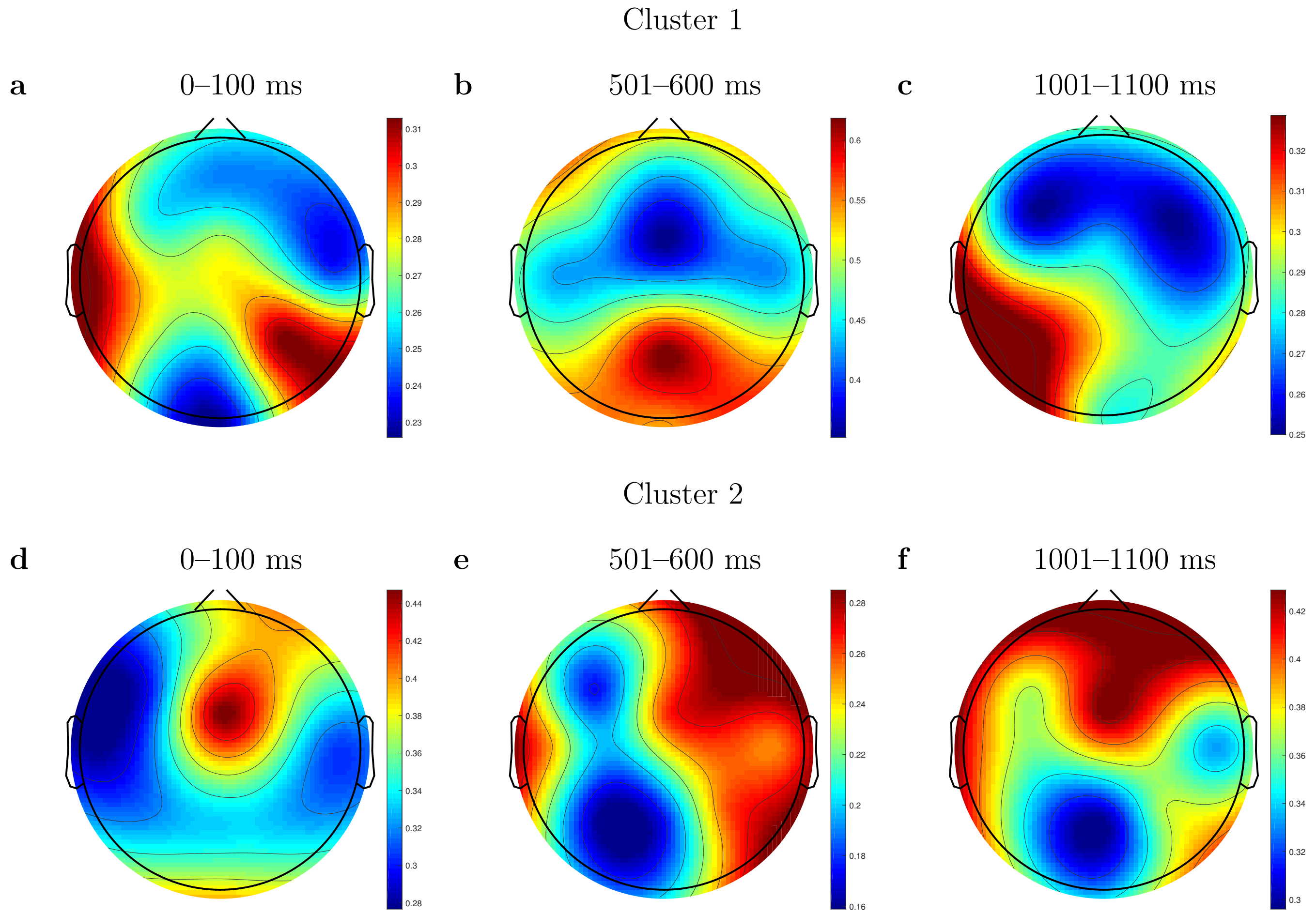

Figure 11. Cluster analysis of feature weights for all participants with LDA AUC > 0.5. A set of two clusters best explained our data ($N = 22$ for cluster 1 and $N = 21$ for cluster 2). (a–c) refers to cluster 1, (d–f) refers to cluster 2. Colors are range scaled. Note that the color scale varies across panels. See Figures 16 and 17 for full version of this figure.

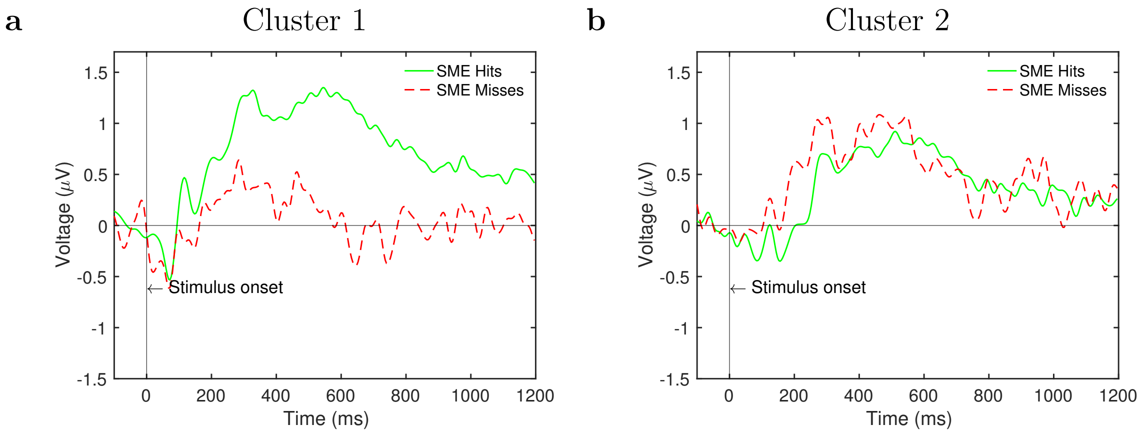

Figure 12. ERPs at Pz for the two clusters obtained through k-means clustering of LDA feature-weights.

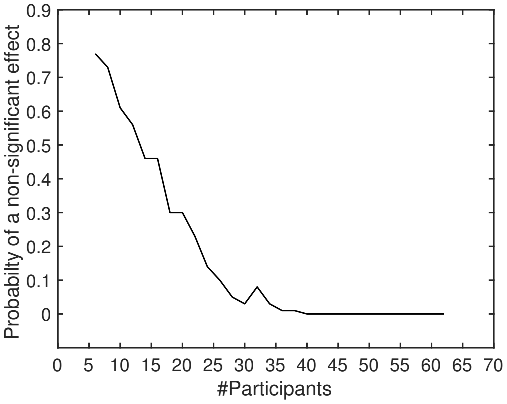

Figure 13. Effect of sample size on the overall significant results for SVM. With one sample t-tests, we calculated if SVM performance was significantly better than chance, for different sample sizes, ranging from 6 to 62 participants. For each sample size, participants were selected at random and without replacement. Further, for each sample size, we collected 100 sets of participants. Y axis shows the probability of obtaining a non-significant effect, calculated across these 100 sets and for each sample size.

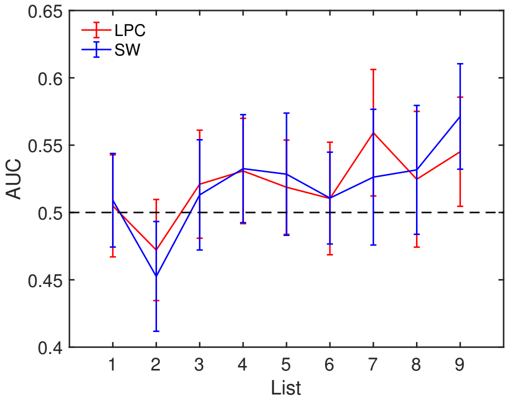

Figure 14. Classification of hit versus miss trials for each list in the task, based on the LPC and SW ERP measures. Error bars are 95% confidence intervals. Dashed line refers to chance performance. Lists with all hits or all misses were excluded.

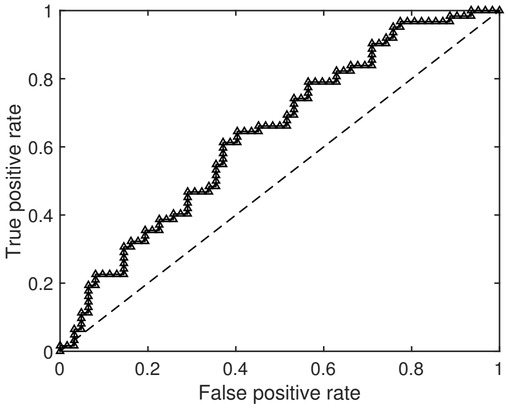

Figure 15. ROC curve obtained from between subject classification of the average EEG waveform at study for hit versus miss events, with linear SVM. Dashed line denotes chance.

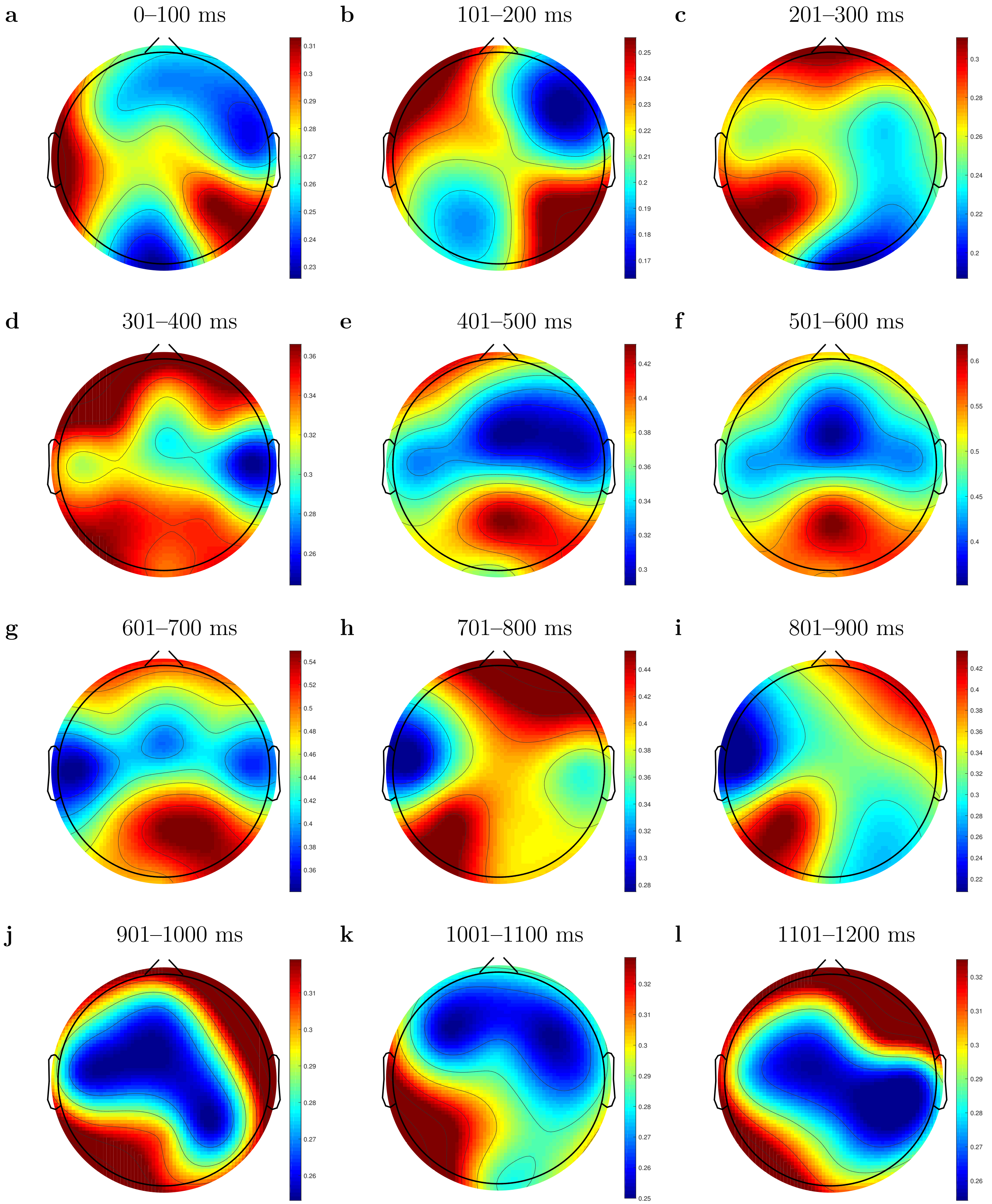

Figure 16. Topographic plots showing LDA feature weights averaged across all participants in cluster 1 ($N = 22$). Colors are range scaled and the scale varies across panels.

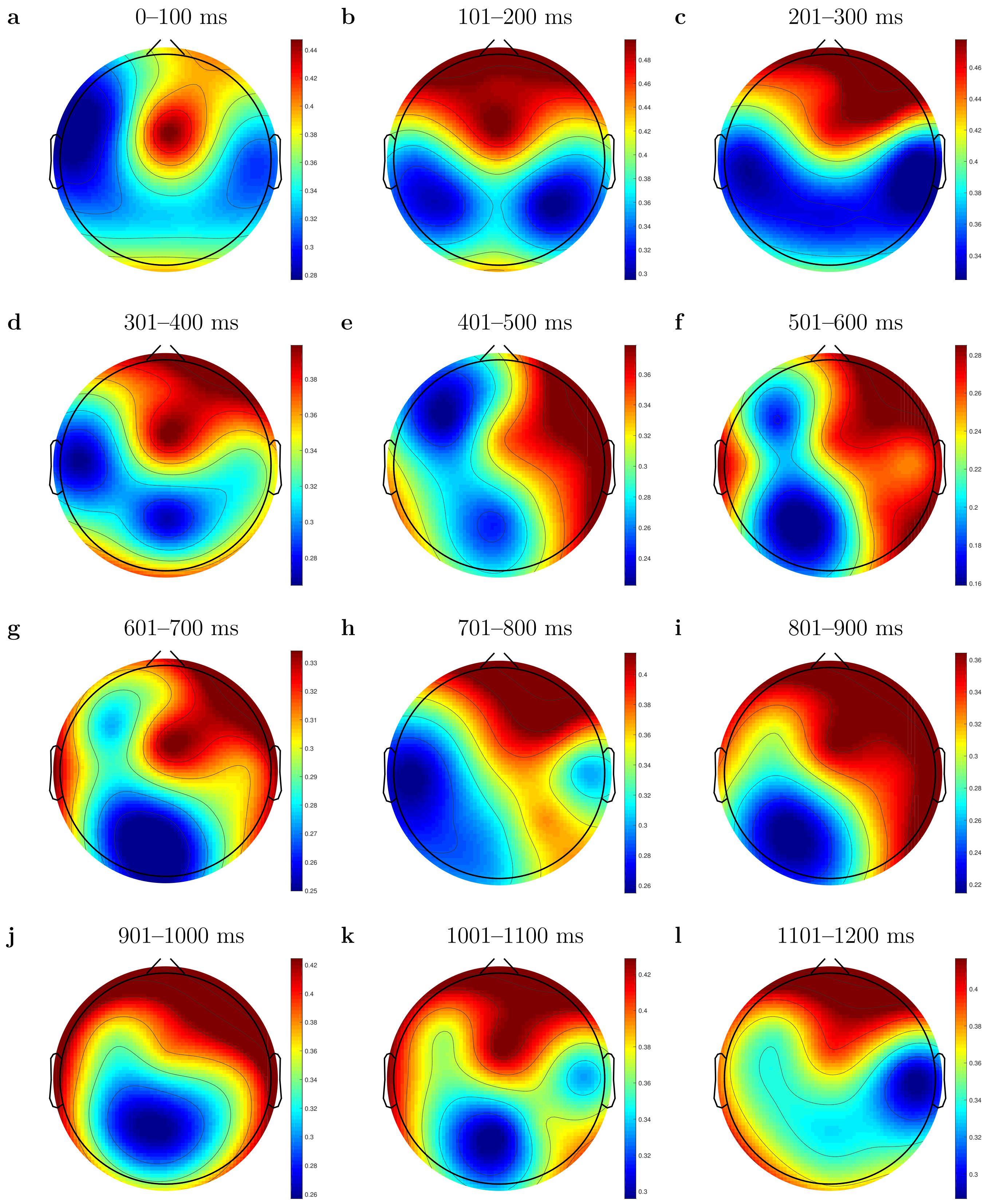

Figure 17. Topographic plots showing LDA feature weights averaged across all participants in cluster 2 ($N = 21$). Colors are range scaled and the scale varies across panels.