EEG-Based Prediction of Successful Memory Formation During Vocabulary Learning

Taeho Kang, Yiyu Chen, Siamac Fazli, and Christian Wallraven

***Abstract*—Previous Electroencephalography (EEG) and neuroimaging studies have found differences between brain signals for subsequently remembered and forgotten items during learning of items - it has even been shown that single trial prediction of memorization success is possible with a few target items. There has been little attempt, however, in validating the findings in an application-oriented context involving longer test spans with realistic learning materials encompassing more items. Hence, the present study investigates subsequent memory prediction within the application context of foreign-vocabulary learning. We employed an off-line, EEG-based paradigm in which Korean participants without prior German language experience learned 900 German words in paired-associate form. Our results using convolutional neural networks optimized for EEG-signal analysis show that above-chance classification is possible in this context allowing us to predict during learning which of the words would be successfully remembered later.**

***Index Terms*—Electroencephalography (EEG), learning, subsequent memory prediction, BCI.**

I. Introduction

HOW the brain encodes, stores, and retrieves memories is one of the core topics across different branches of neuroscience, given its relevance to virtually all aspects of human learning. In the present paper, we investigate the task of studying vocabulary for new language acquisition - a task that requires memorization of words that need to be remembered later. A core question of interest becomes whether it is possible to observe differences in neural activity with respect to those items that will be later remembered correctly versus those that will not. If these differences exist, the neural signatures during the memorization process could be used in a Brain-Computer Interface (BCI) application to alter the studying protocol in order to concentrate on those items that are more likely to be forgotten [1] - an application with vast potential in the realm of language learning. Indeed, such differences have been reported: following Sanquist *et al.*'s seminal study that for the first time observed larger positive neural activities from remembered items than forgotten [2] ones, this difference has also been referred to as subsequent memory effects (SME), or difference due to memory (DM). Since then, neural signatures of SME have been reported in several follow-up studies [3]–[13], although there seems to be a great variability in the details of the signatures: for example, electroencephalography (EEG) studies using pictures or simple words as stimuli with recognition or rating tasks have reported SME as power increases in different frequency-bands including, but not limited to, theta [11], [14], alpha [15], or gamma [16], while other studies reported *decreased power* in such bands, while also reporting increase/decreases in other bands as well [17]–[19]. Hansylmyr and Staudigl provides a more comprehensive review on findings regarding specifics of memory effects [20].

A further complication in the literature arises from the fact that whereas previous studies found SME in post-stimulus signals (i.e., after the picture or word had been shown), other studies also have found significant neural signatures *preceding* stimulus onsets (e.g., [10], along with [4]–[6], [9]. Also see [21]).

A crucial limitation of existing studies on SME is that they often involved measurements of neural signatures contrasting times right before and right after participants were presented with the learning items [4], [6], [9], [11], [16]. Hence, the time between encoding and recall was very short, often mere seconds or minutes after the first presentation. Furthermore, the number of items tested was low in most studies with only a few items presented during the testing phase.

In a typical language learning task, however, the number of items to be studied is usually large and the time between encoding and first test recall is typically in the order of hours or even days. Therefore, in the present study we investigated whether it is possible to find neural signatures of SME in such a more realistic study context.

Among the paradigms suitable for language studying, we chose a *paired associate learning task*, in which participants learn a list of word-pair associations consisting of their native language and a target learning language presented in a flash-card like format. This type of task has been shown to be efficient for vocabulary learning through memorization [22]–[24].

Manuscript received November 28, 2019; revised May 11, 2020 and August 20, 2020; accepted August 24, 2020. Date of publication September 10, 2020; date of current version November 6, 2020. This work was supported in part by the Institute for Information and Communications Technology Promotion (IITP) Grants funded by the Korean Government under Grant 2017-0-00451 and Grant 2019-0-00079, in part by the National Research Foundation of Korea under Grant NRF-2017M3C7A1041824, and in part by the Nazarbayev University Faculty-Development Competitive Research Grants Program under Grant 240919FD3926. *(Corresponding author: Christian Wallraven.)*

Taeho Kang and Yiyu Chen are with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02855, South Korea.

Siamac Fazli is with the Department of Computer Science, Nazarbayev University, 010000 Nur-Sultan, Kazakhstan.

Christian Wallraven is with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02855, South Korea, and also with the Department of Artificial Intelligence, Korea University, Seoul 02855, South Korea (e-mail: wallraven@korea.ac.kr).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2020.3023116

Specifically, our study tested memory prediction by decoding the success of long-term memory formation (i.e., SME) from EEG in the realistic context of prolonged foreign language learning. For this, we gathered neural activity from participants while they spent 5 days learning 900 German-Korean vocabulary word associations without prior knowledge of the German language. Participants were required to recall each word by typing it both on the same day of learning as well as the day after, a manner more akin to realistic studying environments. We also designed the study so that shortly after encountering a new word for the first time, participants would have task blocks where they reviewed and practiced said word several times. With neural activity data gathered from both the presentation period and the ask (practice) period, we investigated *single-trial subsequent memory prediction* by training artificial neural networks with test labels set in two different periods of time after learning.[1] We report that subsequent memory formation can be predicted from stimulus-locked signals collected during learning, as well as feedback-locked signals during practice.

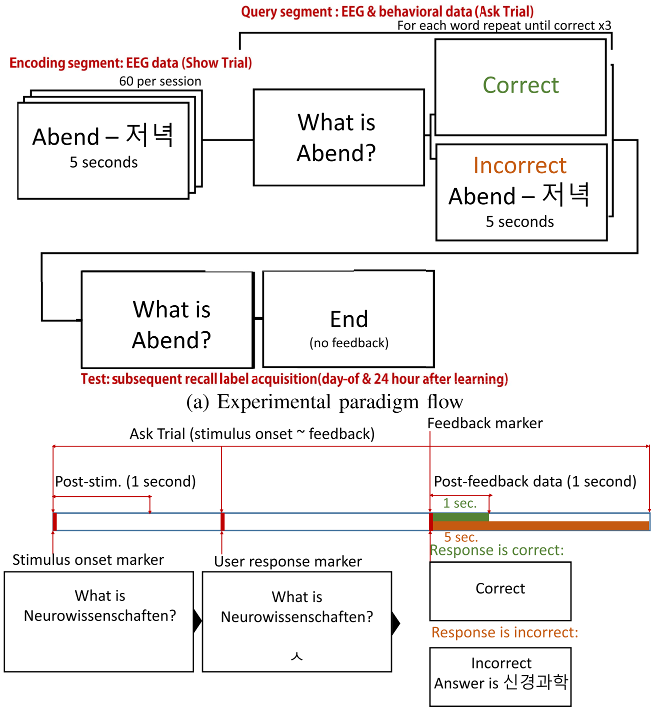
II. Materials and Methods

A. Stimuli

The learning corpus consisted of 2000 German-Korean word pairings that were extracted at random from a German-Korean dictionary of which participants would learn a total of 900 word pairings in order. Due to existing linguistic similarities between German and English and the fact that we could not control for English proficiency in participants, words pairs identical in German and English (e.g. Bus (Ger) - Bus (Eng)) were not included in the stimuli set. Furthermore, participants were encouraged during the experiment to skip word pairs they recognized as familiar due to factors such as similar cognates with English (e.g. Apfel (Ger) - Apple (Eng)). This led to a slightly different list of words being learned for each participant. For a more detailed analysis of linguistic properties of our corpus, please see the supplementary materials.

B. Experiment Procedure

The experimental paradigm and trial definition is shown in Figure 1. Training participants was performed by a two-fold process involving encoding (presentation: from here on we call it "show") and query (review/practice: from here on we call it "ask"). In each encoding segment of a training session, participants encountered new vocabulary by reading a single word-pair juxtaposing Korean and German counterparts that were displayed on screen for 5 seconds. Participants were allowed to skip a word pair and instead learn a different one if they recognized the German spelling as familiar. Initially 10 word pairs were presented in succession, after which the ask segment of the first 10 words began in random order. During each ask segment, participants were shown only the German portion of a word pair on screen and were asked to type the correct Korean counterpart from memory. Once the participant finished their response, feedback was shown. If they gave an incorrect answer, the correct pair was displayed again for 5 seconds. The ask segment for a given word would repeat until it had been correctly answered 3 times, at which point the encoding segment of one new word began. This was followed by resumption of ask segments that included the new word in the pool while excluding the previously correctly-answered word. All trials began with a 1250 millisecond blank interval and ended with a 250 milliseconds of blank interval. In ask trials, between the termination of response and the start of feedback there was also a blank interval of 250 milliseconds.


Query segment : EEG & behavioral data (Ask Trial)
For each word repeat until correct x3
Encoding segment: EEG data (Show Trial)
60 per session
Abend – 저녁
5 seconds
What is Abend?
Correct
Incorrect
Abend – 저녁
5 seconds
What is Abend?
End
(no feedback)
Test: subsequent recall label acquisition(day-of & 24 hour after learning)
Ask Trial (stimulus onset ~ feedback)
Feedback marker
Post-stim. (1 second)
Post-feedback data (1 second)
1 sec.
5 sec.
Response is correct:
Stimulus onset marker
User response marker
What is Neurowissenschaften?
What is Neurowissenschaften?
ㅅ
Correct
Response is incorrect:
Incorrect
Answer is 신경과학

(a) Experimental paradigm flow

(b) Ask trial epoch flow

Fig. 1. Visualization of experimental paradigm and individual trial epochs. Subfigure 1.a shows overall design of a single experiment session. Subfigure 1.b visualizes time point markers within a single ask trial, showing the time points EEG epochs were centered on.

The gist of this design was that at a given ask segment in any point of the experiment, there would be a pool of 10 possible word pairs that could be queried in randomized order. Nearing the end of training sessions in which less than 10 words to learn would remain, extra "distractor" words were added to maintain a pool size of 10 possible word pairs. These distractors were not tested for subsequent long-term memory, and thus were not used for analysis. Note that participants were not informed of the ask pool nor of the existence of distractors.

In each training session, participants learned a total of 60 word pairs, and one day of experiment consisted of 3 training sessions with breaks in between. Shortly after a training session was finished, a short test session consisting of word pairs learned on that day began. Each trial of the test session had a format similar to the ask segment (participants were given the German part of a word pair and had to recall the Korean counterpart), except feedback was not given. Results from test sessions containing word pairs learned on the same day were later used as classification labels for

[1] In supplementary materials, we also investigate whether possible features that may attribute to word difficulty such as corpus frequency and word length have effects on participant memory performance.

same-day memory prediction (the day of learning). Starting from the second day of the experiment, a separate test session would precede the training sessions, testing participants with 180 word pairs learned on the previous day. Results from these separate test sessions were used as classification labels for *next-day* memory prediction (memory performance 24 hours after learning). All participants participated in the experiment for 6 consecutive days to ensure the time between learning and subsequent testing would be in a similar time frame of 24 hours. The experiment paradigm was written with the Pyff framework [25] on Python 2.7.5.

C. Participants

16 university-aged Korean-native participants (all male, age $m = 24.6, s = 1.6$) who had no prior experience with the German language were recruited for the experiment. All participants gave and signed informed consents before proceeding with the experiment. Monetary compensation was provided for participation. One participant did not finish the experiment for personal reasons, and another participant's data was excluded from analysis due to an error in the recording equipment, leaving a total of $n = 14$ participants' data for analysis. The experiment received IRB-approval with number KUIRB-2019-0043-01.

D. EEG Recordings

ActiCAP electrodes and a BrainAmp Amplifier from Brain Products GmbH were used for signal acquisition at a sampling frequency of 1000Hz. 62 channels were selected from the extended 10-5 system [26]: F1,5,z,2,6,9,10, FC5,3,1,z,2,4,6, Fp2, FFT7,8, FT9,7,8,10 FC5,3,1,z,2,4,6, FTT7,8, FCC5,6, T7,8, Cz,3,5,4,6, TTP7,8, TP7,9,8,10, TPP7,8, CP5,3,1,z,2,4,6, P3,5,9,z,4,6,10, PO3,7,z,4,8, O1,Oz,O2. An additional channel for measuring eye movements was defined as EOGv1 and placed under the right eye. Later on, 2 EOG channels were derived from measured channels: EOGh from F9 and F10 for horizontal eye movements, and EOGv from EOGv1 and Fp2 for vertical eye movements. Channels were prepared to sub-20kΩ impedance levels before the beginning of the experiment.

E. Data Preprocessing

Our preprocessing pipeline followed best-practices found in [27] as well as in EEG community knowledge bases [28] to minimize artifacts.

Primary pre-processing and epoching was done using the Fieldtrip toolbox [29] and BBCI toolbox [30]. Acquired raw EEG signals were filtered at a pass band between 1 and 40Hz with a 4th-order Butterworth filter after being down-sampled to 100Hz. Three different sets of epochs were created based on time-lock criteria: stimulus-locked epochs from encoding-segment trials ("show trials"), stimulus-locked epochs from ask segment trials ("stimulus-locked ask trials"), and feedback-locked epochs from ask segment trials ("feedback-locked ask trials"). To ensure each ask trial epoch would only contain signals relevant to the task, ask trials with total lengths from stimulus-onset to feedback-onset of less than 1000ms or more than 15000ms were excluded from the analysis. Furthermore, trials with a response duration (duration of typing) less than 80ms were rejected, as such response times only occurred when participants gave a blank answer. Stimulus-locked show trials were created at $[-1000\ 5000]$ms relative to stimulus onset, while stimulus-locked and feedback-locked ask trials were created at $[-200\ 1000]$ms relative to each marker onset.

Before further artifact rejection, epochs exceeding a threshold amplitude range of 1000μV were excluded from the dataset. Each session's epoched data were decomposed using the runICA implementation [31] of Independent Component Analysis (ICA) available in the Fieldtrip toolbox, visually inspected for components containing excessive EOG, ECG, and muscular movements which were then removed via ICA reprojection. After component rejections, channels Fp2 and EOGv1, and channels F9 and F10 were subtracted to create channels containing vertical and horizontal EOG signals, respectively. Finally, an additional amplitude range-based trial rejection with a threshold of 150μV was applied. Of 900 show trials per participant and approximately 3517 (SD = 361) ask trials per participant collected, our conservative trial rejection pipeline led to rejection rates of 10.3% (SD = 10.8%) for show trials, and 16.4% (SD = 9.7%) for ask trials. EEG dataset is available online.[2] Additional scripts and code can be requested from the authors.

F. Data Analysis - Time-Locked Averages

For time-locked average statistical analysis, topographical maps of significant neural signatures dissociating remembered from forgotten words were created using point-biserial correlation coefficients [32] based on Fourier-transformed power spectrum features. To account for Nyquist frequency and the temporal length of epochs, frequency power values were calculated from integer frequency values starting from 3Hz. Z-scores were calculated from signed squares of point-biserial correlation coefficients based on the signal values and the memory recall label. p-values with zero correlation as the null hypothesis were acquired by means of two-sided z-tests. In calculation of grand-average statistics, inverse-variance weighting under a fixed-effects hierarchical model based on the sufficient statistics approach [33] was used. p-values were corrected for multiple comparisons with Bonferroni-correction [34], divided by the product of number of channels and the number of frequency bins in each epoch.

G. Data Analysis - Classification

Classification was performed within individual participants with three different algorithms: regularized Linear Discriminant Analysis (rLDA) with shrinkage, a convolutional neural network with 3 convolutional layers (3lCNN), and a shallow convolutional network with crops (from here on we refer to this network as ShallowConvNet) adapted from [35]. The ShallowConvNet employed a cropped trial design, in which each trial epoch was subdivided into several sub-trials with non-overlapping time crops before being fed through the network and their prediction results were collated for final

[2] Available at https://osf.io/h634f/ and http://deepbci.korea.ac.kr by request.

TABLE I
INDIVIDUAL PARTICIPANT AND GROUP AVERAGE BEHAVIORAL PERFORMANCE SCORES

Participant	same-day scores (%)	next-day scores (%)
1	80.9	43.7
2	51.2	12.4
3	90.7	60.6
4	84.6	35.9
5	88.8	36.7
6	67.6	22.9
7	67.7	30.6
8	80.8	24.1
9	90.4	47.2
10	79.0	51.6
11	90.1	45.0
12	72.5	44.3
13	61.1	28.9
14	96.2	79.8
mean	78.7	40.3
median	80.8	40.2
std	13.0	17.1

prediction of the trial. A filter of size 40 with kernel of size 3 was used for the temporal convolution layer, and a filter size of 40 with kernel size spanning the entire EEG channel length for the spatial convolution layer. The pooling layer had a kernel size of 2 with a stride of 2. Supercrop size of the network was defined as 70 time points for ask trials, and 350 time points for show trials. Output length of the final convolution layer was set to 1. The 3-layer network had filter sizes of 16, 32, 64 in that order. Each convolutional layer had a kernel size of (5 time points x 3 EEG channels). As we focused on brain signals in the present work, the EOG channels (EOGv, EOGh) were not included as features for training.

Python's scikit library [36] was used for LDA training and classification, in which the shrinkage parameter was automatically chosen according to Schäfer and Strimmer's method [37] based on Ledoit and Wolf's lemma [38]. Pytorch [39] was used for implementations of network classifications. To prevent suboptimal model training from imbalanced training data in classification [40], the dataset was balanced prior to training by undersampling the more numerous class. To ensure generalizability of the trained models, 10% of the dataset were withheld as a test set, and the remainder was split into 10 folds for cross-validation (again balanced across remembered and forgotten classes). Classifications used a total of 5 seconds of post-stimulus epochs from stimulus-locked show trials and 1 second of post-feedback epochs from feedback-locked ask trials. Minimally processed time-series data from EEG channels (EOG channels were excluded) were used at resampled rates of 100Hz. Identity of the words were not considered as factors neither when splitting for training/testing, nor balancing for classes. As such, each ask trial from learning of one word was considered as a separate sample.

III. RESULTS

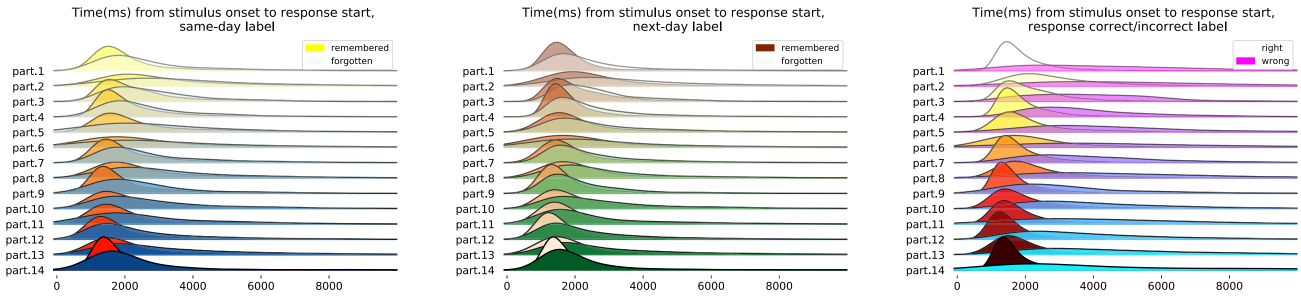
A. Behavioral Analysis

On average, participants correctly recalled 78.7%(±13.0) of the words when tested on the day of learning, while correctly recalling 40.3%(±17.1) 24 hours after learning (see Table I). The difference between performance on same-day and next-day trials was highly significant ($t(13) = 13.16$, $p < 1.0^{-8}$).

A total of 49,244 ask trials extracted from the ask segment were collected across 14 participants. For the present, behavioral analysis, performance from all trials was analyzed (including those ≈10% rejected in the EEG preprocessing). Aside from same-day and next-day labels from subsequent tests that we use to determine successful recall, ask trials provide us with additional label information on whether a given trial was correctly/incorrectly answered during the ask segment. While we did not use correct/incorrect response label information in classification, we use these to perform response-time based behavioral analyses: for example, participants began their response on (Time to response, TTR) median of 1935 (±2742) milliseconds after stimulus onset, and trials where the participant's response to ask queries were correct had significantly faster TTR ($median = 1723 \pm 1930$) than ones with incorrect responses ($median = 3890 \pm 4017 : t(13) = -11.22, p < 1.0^{-10}$). The duration of the response (participants had to type in the word) in ask trials was on average (median) 992 (±1312) milliseconds long. Here, the length of response was longer in trials where the response to the ask trial was correct than incorrect (correct: $median = 1040 \pm 1039$, incorrect: $median = 528 \pm 2034$, $t(13) = 4.21, p < 1.0 \times 1e^{-3}$). We attribute this due to the fact that participants could type in a blank response if they could not think of any possible answer.

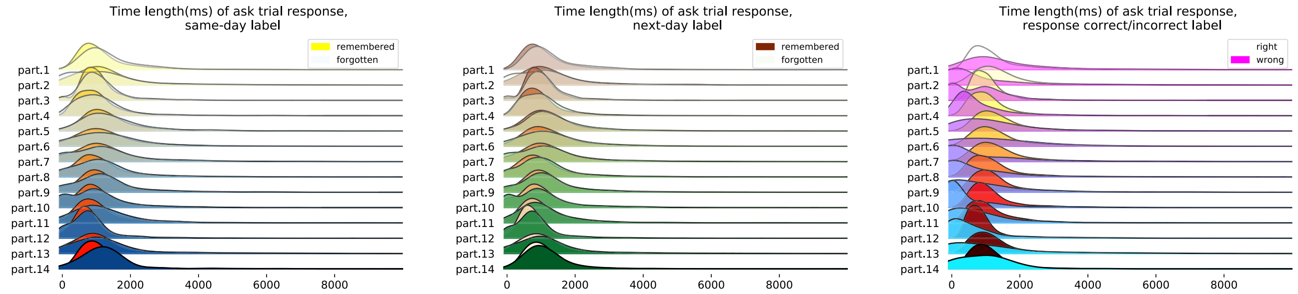
Grouping behavioral response data from ask trials by their subsequent recall labels, we found that TTR was significantly faster on trials with words subsequently recalled on the day of learning (same-day label, remembered $median = 1794 \pm 2349$, forgotten $median = 2529 \pm 3564$, $t(13) = -4.19$, $p < .001$), as well as the day after learning (next-day label, remembered $median = 1597 \pm 1762$, forgotten $median = 2224 \pm 3093$, $t(13) = -4.94$, $p < 1.0^{-4}$). The difference of response duration between subsequently recalled and forgotten information, however, was not significant for both same-day (remembered: $median = 976 \pm 1195$, forgotten: $median = 1040 \pm 1614$, $t(13) = -1.61, p > 0.1$) and next-day (remembered: $median = 928 \pm 1012$, forgotten: $median = 1024 \pm 1450$, $t(13) = -1.57, p > 0.1$).

On average, when a participant subsequently remembered (R) a word in given ask trial (t) on the next day, the probability for the word in the following ask trial ($t + 1$) to have the same label was found to be $P_{next-day}(R_{t+1}|R_t) = 0.39$. In a similar fashion, we explored conditional probabilities of subsequent trial's label given a preceding trial by keeping track of remembered (R) and forgotten (F) class observations: $P_{next-day}(F_{t+1}|R_t) = 0.60$, $P_{next-day}(R_{t+1}|F_t) = 0.36$, $P_{next-day}(F_{t+1}|F_t) = 0.64$. The same analysis was done on subsequent recall using same-day label: $P_{same-day}(R_{t+1}|R_t) = 0.78$, $P_{same-day}(F_{t+1}|R_t) = 0.22$, $P_{same-day}(R_{t+1}|F_t) = 0.74$, $P_{same-day}(F_{t+1}|F_t) = 0.26$. As preceding/subsequent ask trials of a given trial always contained a different word pair, we expected the probability of recall on that given trial to be significantly different from its neighbor trials' probability. From these probabilities, we attempted to test whether recall performance from one trial would influence recall performance of the subsequent trial. If $P(R(t + 1)|R(t))$ and $P(R(t))$ were significantly different, we would interpret this finding as subsequent recall from neighboring trials to being correlated. In other words,



(a) Sameday label (b) Nextday label (c) Correct/Incorrect label

Fig. 2. Within participant kernel density estimation plots of ask trial response time (milliseconds, time from stimulus onset to the start of participant response) by label class. In all label-sets, statistical t-tests between two classes rejected the null hypothesis (same-day: $t(13) = -4.19, p < 0.001$, next-day: $6t(13) = -4.94, p < 1.0 * e^{-}4$, correct/incorrect: $t(13) = -11.22, p < 1.0 * e^{-}10$).



(a) Sameday label (b) Nextday label (c) Correct/Incorrect label

Fig. 3. Within participant kernel density estimation plots of ask trial response time length (milliseconds) by label class. The difference in duration of response between classes were not significant in subsequent recall labels (same-day: $t(13) = -1.61, p > 0.1$, next-day: $t(13) = -1.57, p > 0.1$, correct/incorrect: $t(13) = 4.21, p < 0.001$).

this would be a way to test for the notion whether "streaks" were present in recall performance that are perceived in real-life learning as pockets of time in which the learner is more motivated to learn than usual or vice versa. For example, a t-test performed between $P(R(t+1)|R(t))$ and $P(R(t))$, would test whether the probabilities $R(t)$ [word from a given ask trial would be subsequently remembered] and $R(t+1)|R(t)$ [word from a ask trial would be subsequently remembered, given that word from the preceding trial T1 was remembered] are significantly different. Paired t-tests of the conditional probabilities and their underlying probability (probability of the preceding trial;) rejected the null hypotheses for next-day labels ($t_{R(t+1)|R(t) and R(t)}(13) = 3.52,\ p < 1.0 \times 1e^{-2}$; $t_{R(t+1)|F(t) and F(t)}(13) = -2.93$, $p < 0.05$; $t_{F(t+1)|R(t) and R(t)}(13) = 2.50,\ p < 0.05$; $t_{F(t+1)|F(t) and F(t)}(13) = 2.66,\ p < 0.05$), and mostly rejected for same-day labels ($t_{R(t+1)|R(t) and R(t)}(13) = 1.97$, $p > 0.05$; $t_{R(t+1)|F(t) and F(t)}(13) = 6.02,\ p < 1.0 \times 1e^{-4}$; $t_{F(t+1)|R(t) and R(t)}(13) = -7.65,\ p < 1.0 \times 1e^{-5}$; $t_{F(t+1)|F(t) and F(t)}(13) = 3.98,\ p < 0.01$).

B. Time-Locked Averages

In spectral analysis of stimulus-locked show trials, Bonferroni-corrected statistical tests for next-day labels showed differences between remembered and forgotten items (see Figure 4), but not under same-day label. Specifically, we found significantly lower power in subsequently-remembered items primarily in theta band frequencies as well as some of the alpha band frequencies.

Spectral analysis of the feedback-locked epochs revealed similar results: significant differences between remembered and forgotten conditions were found for next-day labels after applying Bonferroni-corrections, but not for same-day labels. Significance was primarily observed in theta band frequencies, as well as some lower gamma frequencies (see Figure 5). In stimulus-locked ask trial data, spectral analysis of the remembered and forgotten conditions did not yield significant effects in either labels.

C. Classification

Initially, classification was performed with post-stimulus stimulus-locked show trials and post-feedback feedback-locked ask trials, using memory performance 24 hours from learning (next-day) as class labels in assessing whether long-term memory classification was possible at above-chance level. Results from these two classification tasks using three different algorithms can be seen in Table II, along with box-plots of participant-specific results from training with ShallowConvNet, the best performing algorithm, in Figure 6. On average, next-day single-trial memory performance could be predicted using ShallowConvNet with $.540(\pm 076, SD)$ accuracy in post-stimulus show trials, and $.557(\pm 038, SD)$ in post-feedback ask trials.

To compare the above results with chance-level prediction, we ran a simulation of calculating theoretical chance level confidence limits given a data sample size based on [41] using $25,000$ iterations. 10-fold average test accuracy from 7 out of 14 participants' show trial data were higher than upper limits of simulated chance levels at $\alpha = .01$, and

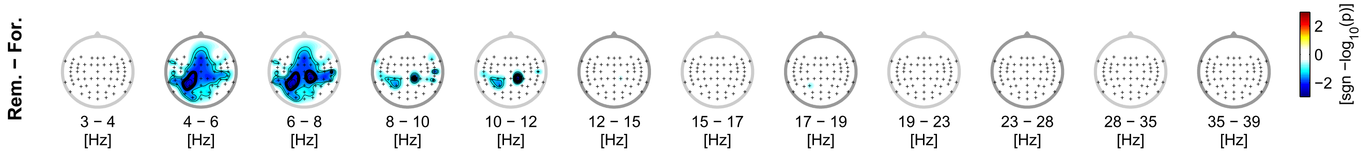


Fig. 4. Scalp-plot of Bonferroni-corrected signed $-log(p)$ values based on Fisher's z scores for show trials using next-day labels for remembered vs. forgotten information in stimulus-locked show trials. Sign of the signed $-log(p)$ values in each data dimension was determined by the difference of average power between remembered and forgotten conditions (negative denotes forgotten conditions had higher values), and p-values were based on Fisher's z scores calculated from point-biserial correlation values [32] of each participant.

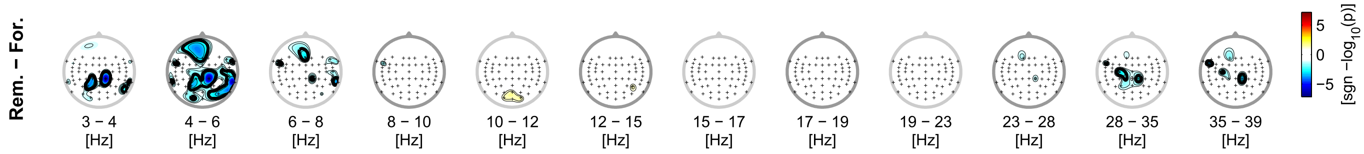


Fig. 5. Scalp-plot of Bonferroni-corrected signed $-log(p)$ values based on Fisher's z scores of frequency data from feedback-locked ask trials using next-day labels for remembered vs. forgotten information. Sign of the signed $-log(p)$ values in each data dimension was determined by the difference of average power between remembered and forgotten conditions (negative denotes forgotten conditions had higher values), and p-values were based on Fisher's z scores calculated from point-biserial correlation values [32] of each participant.

TABLE II

Classification Results of Within-Participant Training. Accuracy Values Are From 10-Fold Averages From the Corresponding Test Set, With Standard Deviation Shown on the Second Row. The Row Labeled "N (Bal.)" Shows the Number of Trials After Balancing for Classes on Each Dataset

5-seconds epochs on stim-locked show trials, next-day label, post-stimulus															
Part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	avg.
LDA	.561 ±.039	.425 ±.090	.560 ±.047	.474 ±.043	.534 ±.028	.367 ±.039	.444 ±.025	.495 ±.038	.484 ±.041	.578 ±.025	.542 ±.024	.511 ±.042	.542 ±.033	.516 ±.059	.502 ±.057
3lCNN	.530 ±.064	.519 ±.112	.531 ±.047	.528 ±.095	.510 ±.074	.550 ±.082	.476 ±.064	.513 ±.069	.543 ±.044	.514 ±.040	.482 ±.044	.506 ±.067	.527 ±.035	.495 ±.075	.516 ±.072
ShallowConvNet	.509 ±.030	.665 ±.045	.575 ±.021	.552 ±.030	.622 ±.035	.494 ±.035	.563 ±.018	.423 ±.026	.596 ±.021	.510 ±.027	.574 ±.028	.615 ±.019	.379 ±.033	.486 ±.044	.540 ±.076
N (Bal.)	622	194	704	536	408	360	550	394	744	872	758	778	518	364	-
1-second epochs on feedback-locked ask trials, next-day label, post-feedback															
LDA	.553 ±.016	.459 ±.020	.574 ±.027	.487 ±.013	.442 ±.022	.471 ±.017	.529 ±.015	.495 ±.034	.491 ±.021	.516 ±.012	.526 ±.017	.529 ±.020	.533 ±.030	.553 ±.036	.511 ±.037
3lCNN	.553 ±.030	.528 ±.064	.541 ±.022	.539 ±.015	.526 ±.037	.552 ±.038	.532 ±.030	.541 ±.033	.558 ±.041	.525 ±.021	.523 ±.023	.520 ±.033	.526 ±.028	.521 ±.047	.535 ±.037
ShallowConvNet	.543 ±.022	.592 ±.043	.602 ±.016	.634 ±.015	.537 ±.026	.563 ±.026	.537 ±.022	.563 ±.023	.570 ±.014	.561 ±.019	.554 ±.022	.551 ±.018	.520 ±.029	.467 ±.044	.557 ±.038
N (Bal.)	2036	558	1998	1750	1412	1146	2014	1310	2628	3164	2464	2328	1916	1140	-

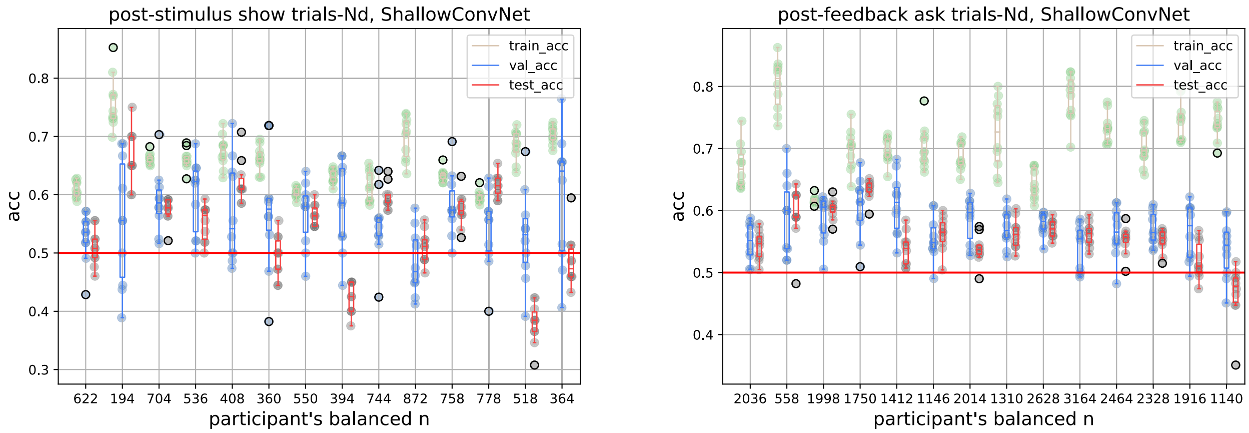
TABLE III

Comparison Between Simulated Chance Levels Extracted From Method [41], With Actual Classifier Performance. Each Simulation Was Run for 25, 000 Iterations With an Alpha of 0.01. The Values Were Compared With 10-Fold Average Test Accuracy From Each Participant

5-seconds epochs on show trials														
Participant	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Actual	.509	.665	.575	.552	.622	.494	.563	.423	.596	.510	.574	.615	.379	.486
Simulated	.553	.588	.548	.556	.564	.568	.553	.563	.546	.544	.546	.545	.556	.566
Above Chance	0	**1**	**1**	0	**1**	0	**1**	0	**1**	0	**1**	**1**	0	0
1-second epochs on feedback-locked ask trials														
Participant	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Actual	.543	.592	.602	.634	.537	.563	.537	.563	.570	.561	.554	.551	.520	.467
Simulated	.528	.556	.530	.530	.534	.539	.529	.536	.525	.523	.526	.527	.529	.539
Above Chance	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	0

12 out 14 participants' results were higher in ask trial data. Individual simulated chance level limits are shown in Table III. As a whole, a t-test between simulated chance and participant average accuracy yielded above chance performance for feedback-locked ask trials ($t(13) = 2.30$, $p < 0.05$), but not for stimulus-locked show trials ($t(13) = -0.792$, $p > 0.05$).

With the best performing network, namely ShallowConvNet, we proceeded to attempt classification on the pre-stimulus (feedback) portion of the epochs as well, along with using labels from memory test performance done shortly after learning (same-day label) instead of the day after (cf.previous studies that reported above chance level prediction on pre-stimulus epochs in shorter-term memory tasks [9], [42]). The results from these runs can be seen in Table IV. Classification results from stimulus-locked ask trials were not included as the classifier did not perform above chance level. With feedback-locked ask trials there were no significant differences in classifier performance between training in post-feedback and in pre-feedback data ($t(13) = 1.942$, $p > 0.05$) nor in next-day and same-day labels ($t(13) = -0.684$, $p > 0.1$). The same



(a) Show trials

(b) Feedback-locked ask trials

Fig. 6. Within-participant classification results from predicting subsequent recall on stimulus-locked post-stimulus show trials (left) and feedback-locked post-feedback ask trials (right) using next-day labels. Each set of box plots show 10-fold cross-validation accuracy results from each participant. Numbers on the X axis denote each participant's class-wise balanced sample size used in classification. The red line denotes the 50% accuracy mark, although it is not necessarily the exact chance level (see Table III). Each dot in the scatter plot represents one data point from a given fold in the cross validation. Dots have borders when data points lie beyond whiskers of the associated box plot. Whiskers are defined as upper (Q3) and lower (Q1) quartiles $+1.5 * IQR(Q3 - Q1)$ of data; any data point that lies outside of this region is labeled as an outlier and therefore had a border.

TABLE IV
Classification Results of Within-Participant Training on Different Trial Epochs and Different Labels. Accuracy Values Are From 10-Fold Averages From Test Set, With Standard Deviation Shown on the Second Row

Participants														
1	2	3	4	5	6	7	8	9	10	11	12	13	14	avg.
pre-feedback epochs on ask trials, next-day label														
525 ±.022	.582 ±.040	.593 ±.019	.570 ±.020	.525 ±.020	.577 ±.038	.532 ±.031	.540 ±.035	.584 ±.017	.511 ±.022	.553 ±.016	.545 ±.020	.494 ±.024	.490 ±.028	.544 ±.032
pre-stimulus epochs on show trials, next-day label														
481 ±.058	.570 ±.056	.511 ±.029	.537 ±.058	.602 ±.049	.511 ±.066	.505 ±.061	.505 ±.046	.536 ±.035	.470 ±.051	.547 ±.044	.582 ±.055	.450 ±.030	.535 ±.051	.525 ±.041
post-feedback epochs on ask trials, same-day label														
542 ±.023	.629 ±.016	.583 ±.035	.527 ±.036	.636 ±.044	.499 ±.018	.502 ±.018	.665 ±.022	.530 ±.039	.556 ±.028	.513 ±.044	.613 ±.021	.594 ±.026	.582 ±.056	.569 ±.051
post-stimulus epochs on show trials, same-day label														
607 ±.034	.628 ±.017	.612 ±.039	.542 ±.062	.400 ±.091	.608 ±.020	.579 ±.051	.497 ±.035	.427 ±.044	.561 ±.046	.394 ±.079	.596 ±.044	.506 ±.024	.657 ±.114	.544 ±.083

was found in stimulus-locked show trials, both between post-stimulus and pre-stimulus ($t(13) = 1.119, p > 0.1$), and between next-day and same-day labels ($t(13) = -0.107$, $p > 0.1$).

D. Classification With Fraction of Chronologically Ordered Full Dataset

Though investigating the feasibility of *on-line* training of memory prediction is not within the immediate scope of this article, we found it prudent to try at least a simulated version of this by training with limited portions of chronologically ordered data. Here we report initial results as an exploration in this direction. For this, we trained multiple ShallowConvNet instances with incrementally increasing portions of the full training data sorted in chronological order, so that the larger training data would contain the latest trials. For each participant we reserved the last 10% of trials from each experimental session as test data. Ten instances of ShallowConvNet were created, each trained with varying portions of the remaining data from 10% to 100% in increments of 10%. From these values, a linear regression of predicting possible accuracy value based on arbitrary portions of the full dataset which may be given as training data was created to see the trend of accuracy values against portion of data used. Based on this regression, we tried to estimate what would be the minimum portion of data to achieve 95% of the regressed accuracy value from using 100% of data, an arbitrarily set value that was deemed adequately comparable to full performance (with chance level of 50% set to be the assumed value when using 0% of data). The corresponding results in Figure 7 show that increasing the number of trials goes along with a general trend of higher prediction performance, although accuracy value comparable to 95% of full dataset performance is not reached on the regression line until above 90% of the dataset is used (black dashed line) in training.

E. Group Level Training With Hold-Out Test Sets

Our classification results up to this point are from training data on a within-participant level. In order to assess whether subsequent long-term memory prediction is generalizable to a larger population, we also trained the network using next-day recall performance as label on a *group-level dataset* validated with leave-one-participant-out cross validation. The dataset was balanced for classes on an individual participant level before combining, in order to guarantee class balance on validation sets. Average test accuracy on group training of feedback-locked ask trials was .520 ($\pm.026, SD$), and .503 ($\pm.020, SD$) on stimulus-locked show trials as can be seen in Table V. While the classification accuracy was lower than within-participant training on average, in feedback-locked

TABLE V

CLASSIFICATION RESULTS FROM TRAINING WITH GROUP PARTICIPANT DATA. ACCURACY VALUES ARE FROM TEST SETS IN A LEAVE-ONE-PARTICIPANT OUT CROSS VALIDATION SCHEME. THE ROW LABELED "TEST SET N" SHOWS THE NUMBER OF TRIALS AFTER BALANCING FOR CLASSES ON EACH DATASET. GROUP AVERAGE TEST ACCURACY IS FOLLOWED BY STANDARD DEVIATION

5-seconds epochs on stim-locked show trials, next-day label, post-stimulus															
Participant (Test set)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	avg.
ShallowConvNet	.490	.521	.510	.547	.490	.494	.482	.515	.536	.490	.513	.479	.483	.495	.503±.020
Test set N (Balanced)	622	194	704	536	408	360	550	394	744	872	758	778	518	364	-
1-second epochs on feedback-locked ask trials, next-day label, post-feedback															
ShallowConvNet	.524	.495	.495	.483	.486	.485	.541	.534	.549	.531	.537	.532	.567	.524	.520±.026
Test set N (Balanced)	2036	558	1998	1750	1412	1146	2014	1310	2628	3164	2464	2328	1916	1140	-

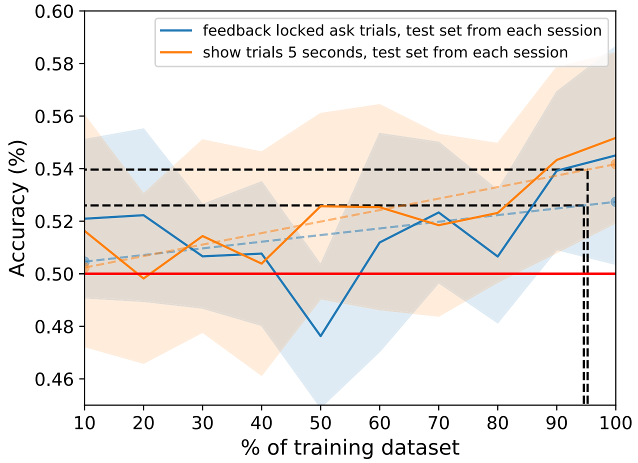


Fig. 7. Result of classification with varying proportions of input data. Plots show mean accuracy levels averaged across participants, as well their 95% confidence interval of the mean, based on what portion of the full dataset was used in training. Colored dashed lines show linear regression of predicting accuracy value given an arbitrary percentage of full data used. Said linear regression was trained on results of 10% increment portions of the full dataset and the resultant accuracy values training the simulated on-line BCI. Black dashed lines represent portion of dataset necessary to reach 95% of the performance of the 100% portion in the regression line, while considering chance level (50%) as the arbitrary 0% portion of the full dataset level.

ask trials it was still significantly higher than simulated chance ($t(13) = 2.871$, $p < 0.05$) - but not for stimulus-locked show trials ($t(13) = 0.573$, $p > 0.1$).

IV. DISCUSSION

In this study, we investigated the possibility of decoding long-term memory formation in a foreign language learning context based on neural data trained on several different classifiers. Based on data collected from participants who underwent prolonged learning of vocabulary of an unfamiliar language, we observed above-chance classification results predicting subsequent memory formation a day from learning. A frequency analysis on time-locked data yielded significant differences between neural data in items tested 24 hours from learning mainly in the theta band, with decreased magnitude on remembered words. Behaviorally we observed that a) memory performance was always worse when tested on a later date, b) participants tended to respond significantly slower in ask trial questions on subsequently-forgotten items, and c) participant recall performance did not correlate with possible factors that may define difficulty of a word such as word length and corpus frequency (see supplementary materials). More specific points for discussion follow.

A. Model Selection and Performance

To thoroughly test EEG signals classification for memory prediction, we decided to choose one established linear model as baseline (LDA), one general convolutional neural network architecture as a representative of nonlinear classification methods, and a more specific network architecture that has been proposed and is intended for EEG dataset classification to see if further performance improvements could be observed. While many different feature extraction methods are available in EEG, for the time being we decided to use minimally preprocessed time-series data only, as the primarily goal of this article was to first see if classification was possible in a practical prolonged memory task. Furthermore, neural networks based classification using minimally processed EEG data without feature extraction has been deemed possible with little performance difference in BCI paradigms [43].

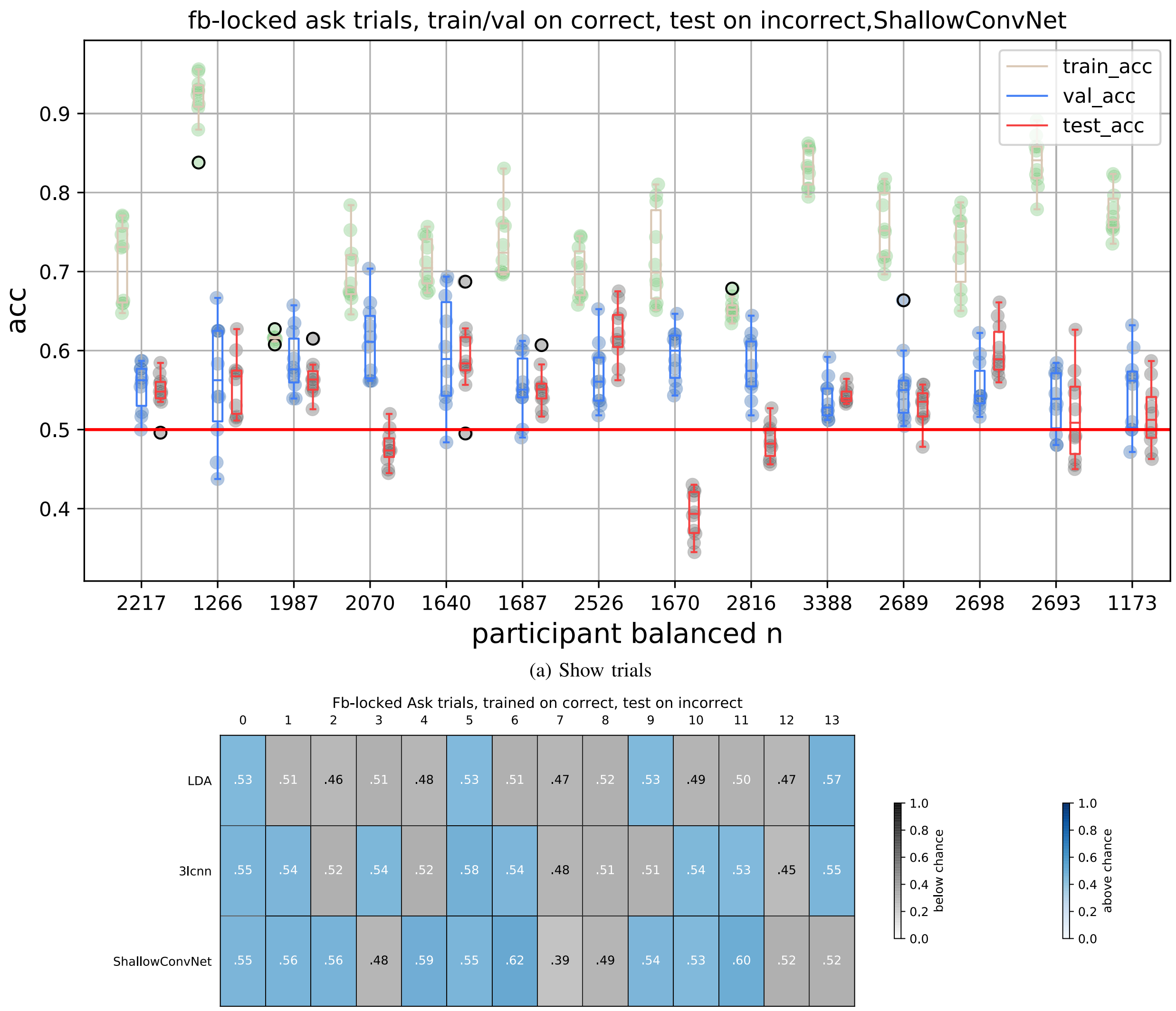
While linear discriminant analysis methods have been known to perform well when trained with established BCI tasks such as motor imagery (MI) and P-300 event related potentials (ERPs), the model performed rather poorly in our dataset. Furthermore, it was outperformed by the Shallow-ConvNet model (Feedback-locked ask trials, $t(13) = 2.70$, $p < 0.05$). This is in contrast to the study where this architecture was introduced which reported comparable performance between regularized LDA and networks trained on MI data [35]. It is unclear why classification could not be reliably made with LDA from our data. One possible explanation may arise from the complexity and general difference in nature of the task, but it may also be due to lack of feature extraction preceding LDA. In Schirrmeister's work of comparing convolutional networks and LDA performance, for example, linear discriminants were not trained on raw time series, but rather on features extracted through processes including temporal band-pass filtering and spatial filtering through Filter-Bank Common Spatial Patterns (FBCSP). Considering that design choices of ShallowConvNet itself were reportedly inspired by FBCSP, it is possible that LDA performance may improve upon such feature selection methods. As our intent was to compare model performances from classifying a task in which feature extraction is not widely applied yet, we decided to forgo explicit feature extraction methods and use unified input data format across all models.

Furthermore, our results indicate that performance from 3-layer CNNs was marginally lower than ShallowConvNet in ask trials ($t(13) = -2.17, 0.01 < p < 0.05$), and not significantly so in show trials ($t(13) = -1.04$, $p > 0.1$). While the better performing network did have more classification results likely to be higher than chance, it is difficult to create conclusions on why performance was higher. Fundamentally, both architectures are forms of convolutional neural networks. The 3lCNN is a bare-minimum implementation of a convolutional network that contains 3 convolutional layers and its design considerations are derived from image processing tasks rather than being optimized for data formats such as EEG. In contrast, ShallowConvNet's layers have specific design choices that emulate a known EEG feature extraction method [35]: the first convolutional layer specifically convolves the time dimension, and the second layer functions acts as a spatial filter, with the kernel encompassing the entire EEG channel size. On top of such parameter choices, it employs a cropped design in which each trial epoch is augmented into smaller crops (although at the end stage crops of each trial merge into one decision). One can assume that these design considerations function as optimizations that improve performance over a basic convolutional network, as seen in our result. CNNs in general can be considered as a feature extractor of their own, and the idea has previously been explored in transferring representations to other classifiers [44]. Perhaps this explains why performance comparable to models trained on extracted features can be observed through training CNNs on raw EEG data, as seen in Fahimi *et al.* [43] In our study, we have used minimally processed data for training across all models; the higher parameter complexity of network models and the resulting depth of feature discovery may be the reason for their superior performance compared to LDA. We believe further investigation by training with previously known features of EEG that are yet to be applied on memory tasks are needed, such as phase-locking values [45] and Hurst exponents [46].

B. On Time-Locked Averaged Results and Classification

In our study, while we observed significant differences between remembered and forgotten items using power spectrum for next-day recall, we were unable to report significant differences between the two conditions using averaged time series (Event related potentials, ERPs. ERP results were not included in this article because of this reason) despite pre-existing literature reporting otherwise [47]. Furthermore, our statistical analysis on power spectrum data did not find significant differences when using same-day label. In many different studies investigating subsequent recall (where stimuli were not necessarily lists of foreign vocabulary but rather native words or even pictures), subsequent memory effects were reported in test sessions performed soon after the end of training session [4], [6], [7], [9], [11], [16], [48]: a time window comparable to our same-day test labels.

We believe there are several possible explanations for the discrepancy with previous findings. First, the stimuli employed in our study were presented in a paired-association form, rather than in a list of words in a single language. Previous EEG studies of memory formation, especially those specifically involving memorization of words, largely had stimuli in the participants' native language, whereas our study had both the participant's native language as well as an unfamiliar language. We believe the presence of a bilingual association pair is relevant: previous psychological findings indicate memorization of verbal-semantic native language association pairs to be more robust to disruptions in short-term memory than memorizing words without association pairs [49]. It should be noted that long-term SMEs were not observed in stimulus-locked ask trials, where only the German word was presented on the screen: in both the feedback portions of ask trials and stimulus-locked show trials where long-term SMEs were found, such association pairs were displayed (in feedback-locked ask trials, the pairs were shown in incorrectly answered trials). Hence, it may be possible that the effects we observed may be correlated with the presence of bilingual association pairs. To test whether displays of association pairs are reflected as features in EEG signal for classification of memory formed, an additional classification was performed: the model was trained and validated with feedback-locked ask trials where correctly-answered responses were made, and then tested on incorrectly-answered trials. As association pairs were displayed in ask trials only if incorrect responses were made (i.e. none in training set), test performance should be worse in this experiment if the displayed association pairs provided an important feature. Test prediction results from this run can be observed in Figure 8. A paired t-test between the previously shown ShallowConvNet test performance from feedback-locked ask trials as seen in Table III and this run did not find significant differences: $t(13) = 0.936$, $p > 0.1$). Given these results, we believe displays of association pairs may not be of primary importance for predicting memory formation. Second, our experiment task involved participants having to accurately recall learned information up to 24 hours after learning, a time-frame much longer than many of the previous studies. We know that memory tends to decay after formation (cf. the decreased next-day performance) so that signal differences between items preserved in memory and items decayed may be more stark if the time of testing is further away from learning. Though our EEG signals in question were collected at the time of learning, if subsequent memory formation is predictable from these signals, it should also be possible that the significant difference along passage of time would be also conveyed in them. Lastly, it is also possible that class imbalance in trials contributed to the lack of significant findings for same-day labels. While only 10.7 words (std: 9.7) on average out of 900 words participants learned were correctly recalled on the following day despite failed recalls on the day of learning, the imbalance between trials is much larger under same-day labels (Remembered : forgotten splits for next-day was 4:6, while 8:2 for same-day label). Because of such imbalance, we believe it may be less likely for statistical tests in same-day prediction to find significant results. Especially in large-dimensional dataset such as EEG, sufficient number of trials are required for both conditions to increase signal-to-noise ratio in comparing binary conditions.



(a) Show trials

Fb-locked Ask trials, trained on correct, test on incorrect

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
LDA	.53	.51	.46	.51	.48	.53	.51	.47	.52	.53	.49	.50	.47	.57
3lcnn	.55	.54	.52	.54	.52	.58	.54	.48	.51	.51	.54	.53	.45	.55
ShallowConvNet	.55	.56	.56	.48	.59	.55	.62	.39	.49	.54	.53	.60	.52	.52

(b) Feedback-locked ask trials

Fig. 8. Within-participant next-day label classification of feedback-locked ask trials, train/validation set from correctly responded trials, test set from incorrectly responded trials. Sub-figure 8a : each set of box plots show 10-fold results from each participant. Red line denotes the 50% accuracy mark, although it is not necessarily the exact chance level. Whiskers are defined as upper (Q3) and lower (Q1) quartiles $+1.5 * IQR(Q3 - Q1)$ of data; any data point that lies outside of this region is labeled as outliers and therefore have a border. Figure 8b : each column represents results from each participant (ordered by participant number), and each row denotes the classifier model used (LDA, 3lCNN, ShallowConvNet). Each cell shows average test accuracy across 10 folds. Cells shaded in blue denote average accuracy exceeding confidence interval tail end of simulated chance levels, while gray denotes not exceeding the tail end of confidence interval.

So far our results indicate little difference in subsequent memory predictability between pre-stimulus (feedback) data trained classifiers and post-stimulus (feedback) data trained classifiers. This is rather puzzling especially in stimulus-locked show trials, where the learning material is not shown to participants until stimulus onset. This result is in line, however, with other studies that observed above-chance classifications using pre-stimulus information: Noh *et al.* [9], for example, observed an increase in prediction accuracy over post-onset epoch classification by combining pre-stimulus data and post-stimulus data. fMRI studies such as Watanabe *et al.* [42] also found above-chance subsequent memory prediction, although leakage of post-stimulus information into pre-stimulus data was suggested as a possible explanation. Given the temporal resolution of EEG this should be less likely to happen, although filtering may cause some dilution of information - consistent with this, we did observe leakage when using Chebyshev type II filters instead of the Butterworth filters employed in the present analysis. Another possible explanation for pre-stimulus prediction is that the underlying, general mental state of the learner is relevant for prediction as well, suggesting that some of the pre-stimulus signal may encode attentiveness and alertness.

Our analysis also showed above-chance classification despite only weak results in the spectral analysis (and no statistically-significant findings for ERPs). This, however, is consistent with observations that univariate statistical methods do not always reveal findings consistent with other methods [50], [51]. It has been suggested, for example, that predictions about underlying brain states made from measured data tend to follow an information-based philosophy with a focus on multivariate decoding methods, while “traditional”

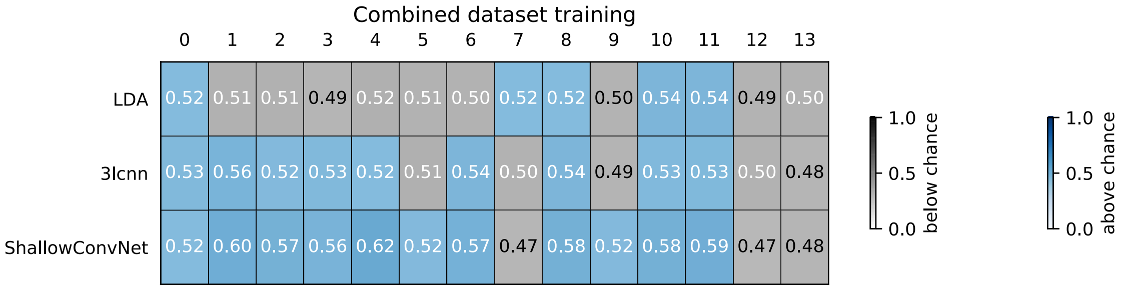


Fig. 9. Within-participant classification results from predicting subsequent recall on the combined dataset of post-feedback ask trials and stimulus-locked show trials using next-day labels. Each column represents results from each participant (ordered by participant number), and each row denotes the classifier model used (LDA, 3lCNN, ShallowConvNet). Each cell shows average test accuracy across 10 folds. Cells shaded in blue denote average accuracy exceeding confidence interval tail end of simulated chance levels, while grey denotes not exceeding the tail end of confidence interval. To fix input size, show trials were "augmented" into 5 non-overlapping 1-second epochs after balancing for classes and splitting. T-test between ask trial only dataset performance and combined dataset performance showed little significant difference in test accuracy performance ($t(13) = 0.804, p > 0.1$).

studies of brain function lean towards a more activation-based philosophy in which univariate statistical frameworks testing linear relationships between data and variables are the norm [51]. Here, predictive methods can reflect nonlinear or non-monotonic effects as well as linear relationships: the classifier may be based on differences in the data distribution, which may not necessarily reflect difference in the mean of neural activity, but possibly variability in measurements. Such information, which may be considered as noise in activation-based philosophy can be considered relevant information in an information-based framework [52].

While we've made speculations on possible reasons for pre-stimulus (feedback) showing predictability, for a better understanding a deeper examination of the trained network should also be considered. Recent developments in the field aiming for explainable artificial intelligence (XAI) [53], [54] has made progress in looking into trained networks through methods such as Layer-wise Relevance Propagation (LRP) [55]. Though the aforementioned method has so far been tested primarily with image classification tasks, we see value in applying this in EEG data as well and believe it is another avenue of future research.

C. Classification on a Combined Dataset

Another question to consider is whether classifier performance could be strengthened by combining data from ask trials and show trials. While signals from the two sets of trials were acquired from different tasks, the learning materials were fundamentally the same. If common predictive features exist, then it would be reasonable to assume training with a combined dataset from the two tasks may see improved classifier performance compared to training with either one of the dataset. To investigate this possibility, we trained new instances of previous within-participant classifier models on combined data, although with certain modifications: because ask trials and show trials were of different time length while our models required static input size, 5-second long show trials were broken up into 5 non-overlapping 1-second epochs. To ensure epochs stemming from the same show trial were not assigned to both training and validation/test sets within folds, this "mini-epoching" of show trials were only done after the dataset was first balanced for classes and split for train/validation/test sets. Furthermore, in order to prevent either ask or show trials from being disproportionately assigned to any one of the sets, train/validation/test set splitting as well as balancing for classes were first done individually for each trial set before merging. With the rest of the pipeline setup remaining unchanged, our results are shown in Figure 9. A paired t-test of test accuracy results from the combined dataset and the ask trial only dataset did not show significant improvement ($t(13) = 0.804, p > 0.1$). While the ShallowConvNet classifier did train for most participants, improvements to the accuracy were minimal despite the greatly augmented sample size. In order to establish whether the learned features are actually shared between the two tasks, a separate run testing for transfer learning by training on one task and testing for the other is needed which is another avenue for future analysis.

D. Application

From a practical BCI standpoint, while the process of memorizing vocabulary is encoded in brain signals potentially much more complex than say, motor imagery, our training results with accuracy levels in the mid-50% range leaves room for improvement. The findings so far indicate that with around half of the full dataset, a comparable level of accuracy can be achieved. We have yet to try more specialized machine learning algorithm in BCI to address issues with dataset size, such as zero-training paradigms [56]–[58]. So far these have been largely validated on well established BCI tasks such as motor-imagery and oddball paradigms rather than more complex, cognitive tasks such as memory formation. While comparable prediction performance could not be reliably achieved with portions of the full dataset in our attempt at simulated on-line learning, we do not believe this eliminates the possibility of applying memory prediction in an on-line context. We have presented initial proof that across-participants classification is possible, albeit with rather limited performance. Considering the above finding, it may be worth investigating whether applying cross-subject transfer learning in simulated on-line BCI of memory prediction could yield better performance, as opposed to the training-from-scratch approach that was used in the present study. Furthermore, EEG signal features are known to vary across time and even sessions [59]. As our attempt at simulated on-line BCI did not account for session-to-session variability of EEG across the dataset, it may also be worth looking into session-to-session transfer as well, although limiting the scope to a session would considerably lower our sample size.

In these contexts, it would be interesting to investigate whether network classifier performance is transferable to other participants or even other tasks, including different memory tasks involving other languages and stimulus modalities. Recently Fahimi *et al.* have explored inter-subject transfer-learning on neural networks trained with EEG data gathered during Stroop task experiments [43], and Thomas *et al.* have shown that across-task deep transfer learning is possible for fMRI data [60]. Given the difficulty in collecting neural data of scale in tasks such as human learning, pre-training classifiers on different tasks may help also in our context.

Taking the relatively low decoding rates of the current approach into account, a viable extension would be multi-modal neuroimaging. Given the low costs, high portability and ease of use, near-infrared spectroscopy (NIRS) would be an obvious addition to the existing setup [61]. Earlier research in BCI has shown that EEG+NIRS measurements do not only significantly increase decoding accuracy of motor imagery, but also that NIRS complements EEG in terms of information content [62], [63]. An obvious concern for including NIRS comes from the long time delay of the hemodynamic response. Thus making the neural decoding within a setup, where association pairs are shown in rapid succession, difficult. However, recently studies have emerged that were able to detect early changes in NIRS patterns and successfully apply these in decoding decisions [64]. Additionally, when focussing on the task at hand, namely predicting the long-term memory formation in single trials one could also explore whether non-neural sources of information, such as eye-tracking, heart-rate, galvanic skin conductance, among others could be beneficial.

V. Conclusion

This study has explored the feasibility of predicting long-term memory in a foreign language learning context. Results indicate that there is a significant difference in power spectrum of signals from learning subsequently-remembered and forgotten information, acquired from when the learning material was initially observed as well as during the ask-segment. Subsequent recall could be predicted at above-chance levels using neural network-based classifiers, but not reliably with linear methods. Initial results showed that participant-tuned online-learning was possible even with a subset of the data and that across-participant generalization in our sample size was achievable as well.

Acknowledgment

The authors would like to thank Benjamin Blankertz for his invaluable ideas, fruitful discussions and help during the design of the experimental paradigm for this study. They would also like to thank Rithwik Mutyala for advice in experimental paradigm and study design.

References

[1] E. Noh, M. V. Mollison, G. Herzmann, T. Curran, and V. R. de Sa, “Towards a passive brain computer interface for improving memory,” in *Proc. 6th Int. Brain-Comput. Interface Conf.*, 2014, pp. 65–68.

[2] T. F. Sanquist, J. W. Rohrbaugh, K. Syndulko, and D. B. Lindsley, “Electrocortical signs of levels of processing: Perceptual analysis and recognition memory,” *Psychophysiology*, vol. 17, no. 6, pp. 568–576, Nov. 1980.

[3] K. A. Paller and A. D. Wagner, “Observing the transformation of experience into memory,” *Trends Cognit. Sci.*, vol. 6, no. 2, pp. 93–102, Feb. 2002.

[4] H. Park and M. D. Rugg, “Prestimulus hippocampal activity predicts later recollection,” *Hippocampus*, vol. 20, no. 1, pp. 24–28, 2010.

[5] J. Fell *et al.*, “Medial temporal Theta/Alpha power enhancement precedes successful memory encoding: Evidence based on intracranial EEG,” *J. Neurosci.*, vol. 31, no. 14, pp. 5392–5397, Apr. 2011.

[6] R. J. Addante, A. J. Watrous, A. P. Yonelinas, A. D. Ekstrom, and C. Ranganath, “Prestimulus theta activity predicts correct source memory retrieval,” *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 26, pp. 10702–10707, 2011.

[7] J. B. Brewer, “Making memories: Brain activity that predicts how well visual experience will be remembered,” *Science*, vol. 281, no. 5380, pp. 1185–1187, Aug. 1998.

[8] A. D. Wagner, “Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity,” *Science*, vol. 281, no. 5380, pp. 1188–1191, Aug. 1998.

[9] E. Noh, G. Herzmann, T. Curran, and V. R. de Sa, “Using single-trial EEG to predict and analyze subsequent memory,” *NeuroImage*, vol. 84, pp. 712–723, Jan. 2014.

[10] L. J. Otten, A. H. Quayle, S. Akram, T. A. Ditewig, and M. D. Rugg, “Brain activity before an event predicts later recollection,” *Nature Neurosci.*, vol. 9, no. 4, p. 489, 2006.

[11] P. B. Sederberg, M. J. Kahana, M. W. Howard, E. J. Donner, and J. R. Madsen, “Theta and gamma oscillations during encoding predict subsequent recall,” *J. Neurosci.*, vol. 23, no. 34, pp. 10809–10814, 2003.

[12] K. Fukuda and G. F. Woodman, “Visual working memory buffers information retrieved from visual long-term memory,” *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 20, pp. 5306–5311, May 2017.

[13] K. Fukuda, M.-S. Kang, and G. F. Woodman, “Distinct neural mechanisms for spatially lateralized and spatially global visual working memory representations,” *J. Neurophysiol.*, vol. 116, no. 4, pp. 1715–1727, Oct. 2016.

[14] W. Klimesch, “Memory processes, brain oscillations and EEG synchronization,” *Int. J. Psychophysiol.*, vol. 24, nos. 1–2, pp. 61–100, Nov. 1996.

[15] P. H. Khader, K. Jost, C. Ranganath, and F. Rösler, “Theta and alpha oscillations during working-memory maintenance predict successful long-term memory encoding,” *Neurosci. Lett.*, vol. 468, no. 3, pp. 339–343, Jan. 2010.

[16] J. F. Burke *et al.*, “Synchronous and asynchronous theta and gamma activity during episodic memory formation,” *J. Neurosci.*, vol. 33, no. 1, pp. 292–304, Jan. 2013.

[17] S. Guderian, B. H. Schott, A. Richardson-Klavehn, and E. Duzel, “Medial temporal theta state before an event predicts episodic encoding success in humans,” *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 13, pp. 5365–5370, Mar. 2009.

[18] J. Fell *et al.*, “Human memory formation is accompanied by rhinal-hippocampal coupling and decoupling,” *Nature Neurosci.*, vol. 4, no. 12, p. 1259, 2001.

[19] T. Staudigl and S. Hanslmayr, “Theta oscillations at encoding mediate the context-dependent nature of human episodic memory,” *Current Biol.*, vol. 23, no. 12, pp. 1101–1106, Jun. 2013.

[20] S. Hanslmayr and T. Staudigl, “How brain oscillations form memories—A processing based perspective on oscillatory subsequent memory effects,” *NeuroImage*, vol. 85, pp. 648–655, 2014.

[21] M. B. Merkow, J. F. Burke, J. M. Stein, and M. J. Kahana, “Prestimulus theta in the human hippocampus predicts subsequent recognition but not recall,” *Hippocampus*, vol. 24, no. 12, pp. 1562–1569, Dec. 2014.

[22] T. Fitzpatrick, I. Al-Qarni, and P. Meara, “Intensive vocabulary learning: A case study,” *Lang. Learn. J.*, vol. 36, no. 2, pp. 239–248, Dec. 2008.

[23] T. Nakata, “Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software,” *Comput. Assist. Lang. Learn.*, vol. 24, no. 1, pp. 17–38, Feb. 2011.

[24] G. Griffin and T. A. Harley, “List learning of second language vocabulary,” *Appl. Psycholinguistics*, vol. 17, no. 4, pp. 443–460, Oct. 1996.

[25] B. Venthur, "PYFF—A pythonic framework for feedback applications and stimulus presentation in neuroscience," *Frontiers Neuroinform.*, vol. 4, p. 179, Oct. 2010.

[26] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution EEG and ERP measurements," *Clin. Neurophysiol.*, vol. 112, no. 4, pp. 713–719, Apr. 2001.

[27] I. Winkler, S. Debener, K.-R. Muller, and M. Tangermann, "On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 4101–4105.

[28] (2019). *Makoto's preprocessing pipeline*. [Online]. Available: https://sccn.ucsd.edu/wiki/Makoto%27s_preprocessing_pipeline

[29] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Comput. Intell. Neurosci.*, vol. 2011, pp. 1–9, Oct. 2011.

[30] B. Blankertz *et al.*, "The berlin brain-computer interface: Progress beyond communication and control," *Frontiers Neurosci.*, vol. 10, p. 530, Nov. 2016.

[31] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski, "Independent component analysis of electroencephalographic data," in *Proc. Adv. Neural Inf. Process. Syst.*, 1996, pp. 145–151.

[32] R. F. Tate, "Correlation between a discrete and a continuous variable. Point-biserial correlation," *Ann. Math. Statist.*, vol. 25, no. 3, pp. 603–607, Sep. 1954.

[33] I. Dowding and S. Haufe, "Powerful statistical inference for nested data using sufficient summary statistics," *Frontiers Human Neurosci.*, vol. 12, p. 103, Mar. 2018.

[34] C. E. Bonferroni, *Teoria Statistica Delle Classi E Calcolo Delle Probabilita*. San Francisco, CA, USA: Libreria Internazionale Seeber, 1936.

[35] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[36] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[37] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Stat. Appl. Genet. Mol. Biol.*, vol. 4, no. 1, Jan. 2005.

[38] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivariate Anal.*, vol. 88, no. 2, pp. 365–411, Feb. 2004.

[39] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. NIPS*, 2017, pp. 1–47.

[40] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.

[41] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller, "Better than random: A closer look on BCI results," *Int. J. Bioelectromagn.*, vol. 10, no. 1, pp. 52–55, 2008.

[42] T. Watanabe, S. Hirose, H. Wada, M. Katsura, J. Chikazoe, K. Jimura, Y. Imai, T. Machida, I. Shirouzu, Y. Miyashita, and Others, "Prediction of subsequent recognition performance using brain activity in the medial temporal lobe," *NeuroImage*, vol. 54, no. 4, pp. 3085–3092, 2011.

[43] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan, "Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based," *J. neural Eng.*, vol. 16, no. 2, 2019, Art. no. 026007.

[44] D. Garcia-Gasulla *et al.*, "On the behavior of convolutional nets for feature extraction," *J. Artif. Intell. Res.*, vol. 61, pp. 563–592, Mar. 2018.

[45] R. Freunberger, W. Klimesch, M. Doppelmayr, and Y. Höller, "Visual p2 component is related to theta phase-locking," *Neurosci. Lett.*, vol. 426, no. 3, pp. 181–186, Oct. 2007.

[46] W. Samek, D. A. J. Blythe, G. Curio, K.-R. Müller, B. Blankertz, and V. V. Nikulin, "Multiscale temporal neural dynamics predict performance in a complex sensorimotor task," *NeuroImage*, vol. 141, pp. 291–303, Nov. 2016.

[47] D. Friedman and R. Johnson, "Event-related potential (ERP) studies of memory encoding and retrieval: A selective review," *Microsc. Res. Techn.*, vol. 51, no. 1, pp. 6–28, 2000.

[48] C. T. Trott, D. Friedman, W. Ritter, M. Fabiani, and J. G. Snodgrass, "Episodic priming and memory for temporal source: Event-related potentials reveal age-related differences in prefrontal functioning," *Psychol. Aging*, vol. 14, no. 3, p. 390, 1999.

[49] C. Papagno, T. Valentine, and A. Baddeley, "Phonological short-term memory and foreign-language vocabulary learning," *J. Memory Lang.*, vol. 30, no. 3, pp. 331–347, Jun. 1991.

[50] M. L. Davidson, "Univariate versus multivariate tests in repeated-measures experiments," *Psychol. Bull.*, vol. 77, no. 6, p. 446, 1972.

[51] M. N. Hebart and C. I. Baker, "Deconstructing multivariate decoding for the study of brain function," *NeuroImage*, vol. 180, pp. 4–18, Oct. 2018.

[52] K. Görgen, M. N. Hebart, C. Allefeld, and J.-D. Haynes, "The same analysis approach: Practical protection against the pitfalls of novel neuroimaging analysis methods," *NeuroImage*, vol. 180, pp. 19–30, Oct. 2018.

[53] D. Gunning, "Explainable artificial intelligence (XAI)," in *Proc. Defense Adv. Res. Projects Agency (DARPA)*, Oct. 2017, p 2.

[54] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 2017, *arXiv:1708.08296*. [Online]. Available: http://arxiv.org/abs/1708.08296

[55] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[56] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, "Towards zero training for brain-computer interfacing," *PLoS ONE*, vol. 3, no. 8, p. e2967, Aug. 2008.

[57] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, "Subject-independent mental state classification in single trials," *Neural Netw.*, vol. 22, no. 9, pp. 1305–1312, Nov. 2009.

[58] P.-J. Kindermans, M. Tangermann, K.-R. Müller, and B. Schrauwen, "Integrating dynamic stopping, transfer learning and language models in an adaptive zero-training ERP speller," *J. Neural Eng.*, vol. 11, no. 3, 2014, Art. no. 035005.

[59] J. M. Krauledat, "Analysis of nonstationarities in EEG signals for improving brain-computer interface performance," Ph.D. dissertation, Fac. IV, Elect. Eng. Comput. Sci., Technische Univ., Berlin, Germany, 2008.

[60] A. W. Thomas, K.-R. Müller, and W. Samek, "Deep transfer learning for whole-brain fMRI analyses," 2019, *arXiv:1907.01953*. [Online]. Available: http://arxiv.org/abs/1907.01953

[61] S. K. Piper *et al.*, "A wearable multi-channel fNIRS system for brain imaging in freely moving subjects," *NeuroImage*, vol. 85, pp. 64–71, Jan. 2014.

[62] S. Fazli *et al.*, "Enhanced performance by a hybrid NIRS–EEG brain computer interface," *NeuroImage*, vol. 59, no. 1, pp. 519–529, Jan. 2012.

[63] S. Fazli, S. Dahne, W. Samek, F. Bieszmann, and K.-R. Muller, "Learning from more than one data source: Data fusion techniques for sensorimotor rhythm-based Brain–Computer interfaces," *Proc. IEEE*, vol. 103, no. 6, pp. 891–906, Jun. 2015.

[64] X. Cui, S. Bray, and A. L. Reiss, "Speeded near infrared spectroscopy (NIRS) response detection," *PLoS ONE*, vol. 5, no. 11, Nov. 2010, Art. no. e15474.