

PAPER • OPEN ACCESS

Continuous decoding of cognitive load from electroencephalography reveals task-general and task-specific correlates

To cite this article: Matthew J Boring *et al* 2020 *J. Neural Eng.* **17** 056016

View the [article online](#) for updates and enhancements.



The Department of Bioengineering at the University of Pittsburgh Swanson School of Engineering invites applications from accomplished individuals with a PhD or equivalent degree in bioengineering, biomedical engineering, or closely related disciplines for an open-rank, tenured/tenure-stream faculty position. We wish to recruit an individual with strong research accomplishments in Translational Bioengineering (i.e., leveraging basic science and engineering knowledge to develop innovative, translatable solutions impacting clinical practice and healthcare), with preference given to research focus on neuro-technologies, imaging, cardiovascular devices, and biomimetic and biorobotic design. It is expected that this individual will complement our current strengths in biomechanics, bioimaging, molecular, cellular, and systems engineering, medical product engineering, neural engineering, and tissue engineering and regenerative medicine. In addition, candidates must be committed to contributing to high quality education of a diverse student body at both the undergraduate and graduate levels.

[CLICK HERE FOR FURTHER DETAILS](#)

To ensure full consideration, applications must be received by June 30, 2019. However, applications will be reviewed as they are received. Early submission is highly encouraged.



PAPER

OPEN ACCESS

RECEIVED
10 February 2020

REVISED
9 September 2020

ACCEPTED FOR PUBLICATION
18 September 2020


PUBLISHED
13 October 2020

Original content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



Continuous decoding of cognitive load from electroencephalography reveals task-general and task-specific correlates

Matthew J Boring^{1,2} , Karl Ridgeway¹, Michael Shvartsman¹ and Tanya R Jonker¹

¹ Facebook Reality Labs, Redmond, WA, United States of America

² Center for Neuroscience, University of Pittsburgh, Pittsburgh, PA, United States of America

E-mail: tanya.jonker@fb.com

Keywords: cognitive load, mental effort, electroencephalography, cross-task, decoding, generalization
Supplementary material for this article is available [online](#)

Abstract

Objective. Algorithms to detect changes in cognitive load using non-invasive biosensors (e.g. electroencephalography (EEG)) have the potential to improve human–computer interactions by adapting systems to an individual’s current information processing capacity, which may enhance performance and mitigate costly errors. However, for algorithms to provide maximal utility, they must be able to detect load across a variety of tasks and contexts. The current study aimed to build models that capture task-general EEG correlates of cognitive load, which would allow for load detection across variable task contexts. **Approach.** Sliding-window support vector machines (SVM) were trained to predict periods of high versus low cognitive load across three cognitively and perceptually distinct tasks: n-back, mental arithmetic, and multi-object tracking. To determine how well these SVMs could generalize to novel tasks, they were trained on data from two of the three tasks and evaluated on the held-out task. Additionally, to better understand task-general and task-specific correlates of cognitive load, a set of models were trained on subsets of EEG frequency features. **Main results.** Models achieved reliable performance in classifying periods of high versus low cognitive load both within and across tasks, demonstrating their generalizability. Furthermore, continuous model outputs correlated with subtle differences in self-reported mental effort and they captured predicted changes in load within individual trials of each task. Additionally, alpha or beta frequency features achieved reliable within- and cross-task performance, suggesting that activity in these frequency bands capture task-general signatures of cognitive load. In contrast, delta and theta frequency features performed considerably worse than the full cross-task models, suggesting that delta and theta activity may be reflective of task-specific differences across cognitive load conditions. **Significance.** EEG data contains task-general signatures of cognitive load. Sliding-window SVMs can capture these signatures and continuously detect load across multiple task contexts.

1. Introduction

As tasks, goals, and the environment change throughout daily life, the cognitive demands on people fluctuate, leading to varying levels of cognitive load [1, 2]. When people are under high cognitive load, they are more likely to make mistakes. These mistakes are often accompanied by small costs, such as lost time or missed opportunities, but they can sometimes result in disastrous outcomes, such as automobile or

aviation accidents [3–5]. As such, there is considerable practical benefit to building systems that can detect cognitive load because they can be used to help people adapt to the cognitive demands of their environment. For example, computers could alert users or automate parts of a task during periods of high cognitive load, which may enhance human performance and mitigate errors.

Physiological signals, such as electroencephalography (EEG), have proven useful for detecting

changes in cognitive load [6–15]. Several studies have developed highly accurate models to classify high versus low cognitive load trials within individual tasks [9, 14, 16–19]. For example, experimenters often use n-back tasks to manipulate cognitive load by varying the number of consecutive trials that must be remembered [10, 11, 13, 20–22]. Studies have then applied machine learning to EEG event-related potentials and/or time-locked spectral features to successfully discriminate between high- versus low-load trials of n-back [11, 13, 14]. Similar approaches have been applied to other tasks as well, including those that vary the complexity of manipulations on items in working memory, such as mental arithmetic [12, 17, 23], or tasks that vary the number of objects or tasks that one must attend to, such as multi-object tracking (MOT) [8, 16, 24].

Although these demonstrations are compelling, it is difficult to determine whether the resulting models can generalize to new tasks. Do the discriminative features captured by these models reflect task-general correlates of cognitive load or do they reflect differences between conditions that are idiosyncratic to each task? For example, increases in cognitive load during n-back tasks are commonly validated by observing increases in response times [17, 25, 26]. As such, EEG components associated with motor preparation and the execution of responses during the task will have different time-courses for high- versus low-load conditions. This can lead to inflated performance estimates of cognitive load models because the model may learn to discriminate based on the timing of motor-related EEG features [27], which would not generalize to tasks that require different types of responses. Therefore, to develop predictive models that capture task-general correlates of cognitive load, it is important to train and validate those models across tasks that vary along perceptual, motor, temporal, and cognitive dimensions. Models that can reliably predict cognitive load across tasks that vary along these dimensions are more likely to be driven by task-general features of cognitive load.

Few studies have attempted to design models that can predict cognitive load across multiple tasks [10, 16, 17, 26, 28] and even fewer have succeeded in achieving reliable performance on novel tasks that were not in the training set [26, 28]. Within studies that have succeeded, it is unclear whether reliable cross-task cognitive load prediction was influenced by confounding differences in motor responses across high- and low-load versions of each task [26], and it is impossible to determine which patterns of EEG sensor activity is associated with changes in load rather than with other task features [28]. Furthermore, current predictive models of cognitive load often focus on classifying *discrete* trials into *binary*—high or low cognitive load—states [9, 10, 16–19, 26, 28], despite evidence that load is dynamic

[14] and has varying levels across tasks [2, 29, 30]. And finally, because current state-of-the-art classifiers are often trained on discrete trials, they cannot be applied to tasks with different trial lengths or task structure. This means that they are also unable to identify changes in cognitive load that occur within trials of different experimental tasks.

The goal of the present study was to overcome these three limitations of previous models used to predict cognitive load across different tasks using EEG data. Specifically, we sought to (1) use a task design that minimized confounds between high- versus low-load trials that were shared across tasks, (2) train models that operate continuously in time, and (3) provide a continuous (rather than binary) index of cognitive load. We chose to manipulate cognitive load using variants of the n-back, mental arithmetic, and MOT tasks, which allowed us to capture changes in cognitive load that are induced by different component cognitive processes, including updating and maintaining working memory, manipulating working memory, and splitting endogenous attention, respectively [22, 31, 32]. The tasks were designed such that the low-level visual features, the temporal structure of the trials, and the motor responses varied across tasks. By varying these perceptual, temporal, and cognitive features across tasks while manipulating load within each task, we could determine whether the classifiers could learn task-general correlates of high versus low cognitive load. Furthermore, by characterizing the time-course of model outputs over trials, we could provide insight into the time-course of cognitive load fluctuations within each task.

We hypothesized that across n-back, mental arithmetic, and MOT tasks, increases in cognitive load would be reflected in task-general changes in spatio-spectral features in the EEG data. Sliding-window support vector machines (SVMs) were used to attempt to learn these features. Based on previous studies, we predicted that changes in theta and alpha power should be predictive of cognitive load across tasks [13, 14, 20, 26]. Furthermore, we hypothesized that cognitive load would evolve with distinct time-courses across the three tasks. Specifically, we predicted that (1) the n-back task would produce sustained levels of load because working memory is constantly maintained and updated on each trial, (2) the mental arithmetic task would produce transient increases because the complexity of numerical manipulation increases throughout a trial, and (3) the MOT task would produce sustained levels of load during the tracking phase because exogenous attention is split between a constant number of visual targets. Finally, to test if the models were sensitive to differing levels of cognitive load within high- and low-load conditions, we correlated model outputs with self-reported mental effort.

2. Methods

2.1. Participants

EEG data from 33 participants were analyzed for the current study. These participants were part of a larger group of 67 participants that participated in a single 3-h experimental session where pupillometry, eye-tracking, galvanic skin response, and heartrate were also collected (these data streams were collected for use in a different study and are not analyzed here).

Of the total 67 participants, 34 participants were excluded for various reasons related to the technical complexity of the study: 14 failed to complete the experiment due to participant non-compliance or loss of data quality during the recording session; three participants completed the session without EEG data collection; ten participants' data were lost due to technical disruptions between the stimulus presentation and recording software which resulted in a loss of stimulus timing information; and seven participants were excluded from analysis due to excessive environmental noise and large DC offsets in the EEG data indicating poor electrode contact. The remaining 33 participants had a mean age \pm SD of 27.6 ± 5.7 years, ten females, five left-handed.

Informed consent was obtained, and all experimental protocols were approved by the Western Institutional Review Board.

2.2. Experimental paradigm

During the session, participants completed a resting phase, followed by three tasks in the following order: n-back, mental arithmetic, and MOT (described below). The tasks were designed to manipulate cognitive load while varying their perceptual, temporal, and cognitive features to help disentangle differences in cognitive load from potential confounding features (e.g. response time differences) across high- and low-load trials within each task (see figure 1 for schematics of each task, including stimulus presentation times). Prior to the study, we ran a pilot study with ten participants to ensure that high- and low-load blocks of each task evoked similar subjective reports of mental effort according to the NASA-Task Load Index (TLX) [29].

Before the first block of each task, participants completed a practice session, including trials of both the high- and low-load conditions. Visual instructions were presented at the beginning of each block to tell participants whether to perform the high- or low-load version of each task. Blocks within tasks were separated by self-paced breaks. Tasks were separated by self-paced breaks of at least 5 min. The sequence of the tasks and the order of high/low blocks within each task were held constant across all participants.

2.2.1. Resting phase

During the resting phase, participants were asked to relax and let their mind wander while fixating on a

central fixation cross. We collected 5 min of resting state data, which were later used for baselining the EEG data.

2.2.2. N-Back

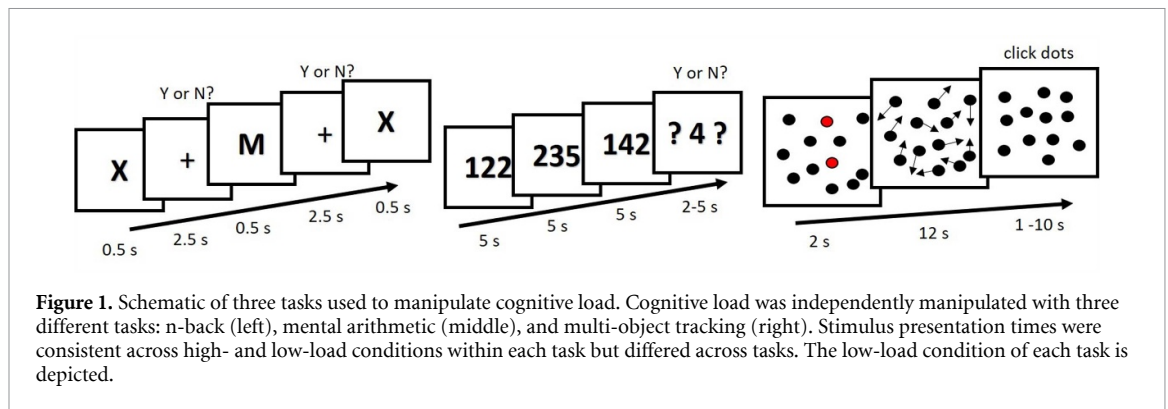
After the rest phase, participants completed 10 blocks of 0-back and 2-back tasks (5 each) in a pseudo-random order that was consistent across participants (2-0-0-2-0-2-0-2-2-0). At the beginning of each block, participants were informed of the block's difficulty (i.e. 0-back vs. 2-back). Next, a series of 40 consonants were presented individually for 0.5 s each with 2.5 s between each letter. For 0-back trials (low cognitive load), the participant was instructed to press the right arrow key on the keyboard if an 'x' was presented on the screen and the left arrow key if any other letter presented. The letter 'x' was presented ten times during each block of the n-back task (25%). For 2-back trials (high cognitive load), participants were asked to press the right arrow key every time they saw a letter that matched the one presented two trials before and the left arrow key if it did not. Each block contained ten trials where the letter was repeated two trials prior (25%). Across all blocks, there were 200 3 s trials for each condition per participant.

2.2.3. Mental arithmetic

After the n-back task, participants completed ten blocks of single- or triple-digit mental arithmetic trials (five each) in a pseudo-random order that was consistent across participants (3-1-1-3-1-3-1-3-3-1). Each block contained six trials. For the high-load condition, participants were sequentially presented with three 3-digit numbers, which they were to sum. They were then shown a fourth number and were to indicate whether it equaled the sum of the three 3-digit numbers. The final number was different from the true sum 50% of the time by either ten or one. Each number in the sum was presented for 5 s and participants were given up to 5 s to indicate their response (left arrow if correct, right arrow if incorrect) at the end of the trial. For the low-load condition, participants were also shown three 3-digit numbers; however, they were instructed to only add up the leftmost digit of each number. The proposed sum of low-load trials was off by one 50% of the time. There were 30 trials for each condition per participant.

2.2.4. Multi-object tracking (MOT)

Participants completed ten blocks of 2-dot and 6-dot MOT tasks (five each) in a pseudo-random order that was consistent across participants (6-2-2-6-2-6-2-6-6-2). For the high-load condition, participants were presented with a random array of 14 blue dots. At the start of each trial, six of the dots flashed red for 2 s. Participants were instructed to mentally track the motion of these six dots. After 12 s of random dot motion, participants were prompted to click on the six target dots. They had up to 10 s to respond. For



the low-load condition, all trial features were similar, except that participants were shown only two dots as targets. There were 30 trials for each condition per participant.

2.2.5. Surveys

At the end of each block for all tasks, participants answered a series of questions, including the NASA-TLX [29] to assess the experienced difficulty of the blocks of trials. They were then shown their accuracy for the previous block of trials to increase their motivation to engage in the tasks.

2.3. Data collection and preprocessing

EEG data were collected using a Biosemi ActiCap 64-channel system (Biosemi B.V., Amsterdam), with two additional channels, one placed on each mastoid. Electrodes were placed according to the standard 10–20 system and data were collected at 512 Hz. During recording, data were visually inspected for clear alpha activity and blink-related artifacts to ensure signal quality. Data from 33 participants were preprocessed and used for the cognitive load classification analyses presented here.

EEG data was preprocessed using the EEGLAB MATLAB toolbox (ver. 2019.0) [33] (figure 2). First, data were referenced to a common average of the two mastoid electrodes, bandpass-filtered from 0.5 Hz to 50 Hz, then down-sampled to 250 Hz. Noisy channels were defined as those that were not correlated with neighboring channels ($r < 0.8$) or that flat-lined for longer than 5 s at some point during the experiment. These channels were removed from the dataset and their data were estimated from neighboring channels via spherical interpolation. Next, EEG data were cleansed using artifact subspace reconstruction (burst criterion = 20), which automatically detects and removes high amplitude artifacts arising from eye-blinks, muscle contractions, or movement using a sliding PCA approach [34]. This algorithm has been demonstrated to effectively remove motion and muscular artifacts while preserving underlying neural activity [35]. After artifact subspace reconstruction, we removed periods of time during which EEG data from more than 50% of channels were marked as bad

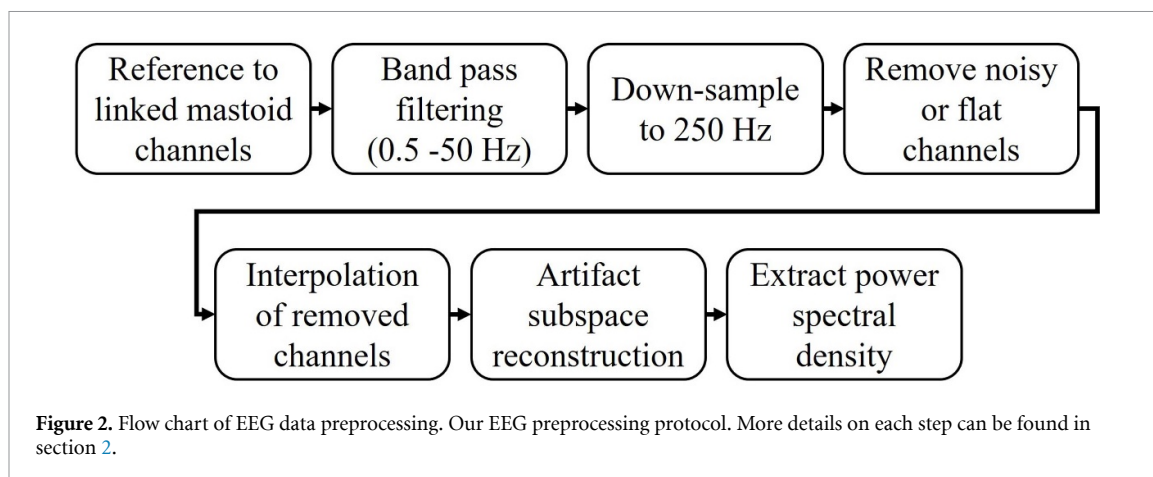
due to signal decorrelation induced by voltage jumps or sensor movement.

Preprocessed EEG data were then decomposed into spectral features, i.e. power spectral density, using short-time Fourier transforms (STFT). One-second Hann windows were used to calculate the STFT to balance accurate estimation of signal power in low EEG frequency bands, while maintaining fair temporal resolution of our estimates [36]. A half-second stride length between STFT calculations was chosen to balance the temporal resolution of successive power estimates, while not oversampling the EEG time-course (since successive STFT Hann windows overlapped by 0.5 s). Oversampling the EEG time course would increase the computational demands of subsequently training and testing classifiers because there would be more time windows to analyze.

This procedure yielded 45 spectral features (1–45 Hz) at 64 EEG channels for each 0.5 s of the experiment across the 33 participants with clean EEG data. We chose to use time-frequency EEG features because they capture changes in neural oscillations that are more phasic than voltage potentials (depending on the length of the window used to calculate the Fourier transform) and do not rely on clear stimulus onset information like traditional ERP analysis [13]. As such, time-frequency features are more promising for learning EEG correlates of cognitive load fluctuations in the real world, where explicit knowledge of stimulus timing is often unknown.

2.4. Sliding-window models for continuous cognitive load prediction

The goal of this study was to develop models that can index cognitive load from EEG data across a variety of cognitively, temporally, and perceptually diverse tasks. To explore the space of predictive models thoroughly, we opted to explore the performance of several commonly used classification algorithms (SVMs, k-nearest neighbors (kNN), and linear discriminant analysis (LDA)) and different choices of parameters for each algorithm. We chose to use a sliding-window approach, treating time-windows of data across high- and low-load blocks as independent ‘trials’—rather than the trials themselves—to train these algorithms.



This sliding-window approach allowed us to apply the same classifier to tasks with different trial lengths because the start of the window was not locked to stimulus onsets and was not designed to capture the full time-course of a trial. We then slid this fixed window along the length of the trial (stride = one time sample (0.5 s) and made predictions of the level of load within our selected window size. We chose to use this stride length because a longer stride would decrease the number of training samples (since they would be further apart in time) and would decrease the temporal resolution of our predictions (since they also would be further apart in time). The length of the sliding window was a less clear choice. Therefore, we tested several window sizes (see below). Applying the trained classifiers on sliding-windows of test data produced a continuous readout of load (every 0.5 s) throughout the course of the experiment.

SVMs with gaussian radial basis function kernels were trained using the LIBSVM toolbox [37]. Specifically, SVMs were trained to use spectral features across all 64 EEG channels within a given time window to predict whether a participant was completing a high- or low-load trial during that time window. We chose these types of models for our problem of cognitive load prediction because they are powerful in regimes where the number of features is larger than the number of training instances and they do not require as much tuning as other non-linear models, such as neural networks [38–40].

Labels for high- versus low-load classification were coded as a vector indicating whether each time point of the experiment belonged to a high-load trial, a low-load trial or an inter-trial interval. For n-back blocks, we hypothesized that load would be sustained throughout the course of a block because participants were constantly updating their working memory and comparing stimuli to its contents to complete the task. Therefore, we labeled all time points from the first to last trial of the block as high- or low-load. Because mental arithmetic blocks contained six independent trials, we labeled all time points from the

first integer of each trial to the key-press response as high- or low-load. For MOT, we labeled only time points corresponding to the 12 s tracking phase as either high- or low-load. All remaining time points were considered inter-trial interval. During training and testing of the sliding-window classifiers, the label of a given time-window corresponded to the most prevalent label (i.e. the mode) within that window. Both correct and incorrect trials were included in the analysis. Time windows labeled as inter-trial interval were not used in training or testing.

All models were trained and tested on data independently for each participant. Within-task SVM classifiers were trained and tested on data from a single task (n-back, mental arithmetic, or MOT) and underwent fivefold cross validation, where each of the five partitions contained data from one high- and one low-load block. This ensured balanced samples and that overlapping time windows from the same block were not included in both train and test data. Cross-task models used the same spectral EEG features as within-task models for classification, but they followed a different cross-validation procedure. Cross-task SVMs were trained using all blocks of data from two of the three tasks, and they were tested on all blocks from the held-out task. This procedure was performed independently for each of the three tasks. The default C and SVM parameters of 1 and $1/\text{number of features}$ (64 channels \times 45 frequencies \times 6 time points for the main analyses) were used because they achieved reliable performance and optimizing these parameters comes at a large computational cost. If optimized, we would expect slight improvements to our SVM model performances.

We also tested the performance of alternative models for sliding-window cognitive load classification. We chose to compare SVMs with LDA and kNN algorithms due to their popularity and prevalence in EEG signal classification [10, 13, 26, 40]. These algorithms were implemented using built-in MATLAB functions. For kNN we explored three different choices of k (3, 5, and 9), which defines the number

of neighboring examples that are used to classify each test example. All models were trained using the same cross-validation strategy described above.

Finally, we also compared performance of SVMs trained with different window lengths. Longer window sizes may provide higher predictive performance because the classifiers may be more robust to noise at individual time-points. However, these longer windows are not ideal for interpreting the continuous time-course of model outputs because it is unknown which time points within these longer windows contributed to the model's prediction. Ultimately, the size of the sliding window is a practical decision. To provide a complete picture of the data, we report a comparison of SVM classifiers trained on 1-s, 3-s, 6-s, and 12-s windows. To compare the effects of increasing the classifier's time window without confounding the effect of adding more parameters to the classifiers, we compared the performance of 3-s window classifiers to 6-s and 12-s window classifiers that were all trained on the same number of features. This was achieved by averaging over 2 or 4 consecutive time points for the 6-s and 12-s window classifiers, respectively. Decreasing the number of features of the 3-s window classifier to equal the number of the features of the 1-s long classifier did not improve classification performance (data not shown). We focused all of our in-depth comparative analyses using 3-s sliding-windows (six temporal features, 1 sample stride length) because that was the length of our shortest type of trial (n-back) and achieved almost the same performance as classifiers trained on longer time windows.

2.5. Statistics

2.5.1. Behavioral analyses

To analyze behavioral data, we used mixed-effects models [41], as our measurements were repeated within participants. In all models, our objective was to estimate covariates (e.g. of block and trial) over all tasks simultaneously and control for the random effect of participant. In all models, we estimated a maximal random-effect structure (i.e. with random slopes and intercepts for within-participant parameters of interest but only random intercepts otherwise). We did not estimate random slopes for covariates. We used zero-sum contrasts to allow for direct comparison between levels of a variable; for example, the effect size of load is directly interpretable as the difference between high- and low-load. Task was coded using treatment contrasts with the n-back task as baseline, given that the n-back task is often treated as the 'gold-standard' task for working memory and cognitive load [42].

Note that for a balanced design (such as in our analysis of the effect of cognitive load on NASA-TLX scores), the coefficients of mixed-effect models are identical to the corresponding linear model

coefficients, but the standard errors tend to be larger (more conservative) to better account for correlations between observations for a single participant. The analyses generalized paired t-tests or repeated-measures ANOVA because we were able to simultaneously control for covariates (e.g. practice effects due to the duration of the experiment) and for continuous covariates like trial-wise and block-wise practice effects. To ensure that the models converged with full random effects structures, we used a boundary-avoiding prior on the random effects covariance [43] using the *blmer* package for R.

For behavioral statistical models, we report bootstrapped 95% confidence intervals. Informally, confidence intervals are a statement about the reliability or replicability of our estimates: If our experiments were repeated many times, the fraction of estimates in these other experiments that fall within the intervals we report would approach 95%. We do not report *p*-values for these analyses, as correct degrees of freedom for mixed-effects models are not known except in some special cases.¹ Readers who prefer to reason in terms of binary significance decisions can interpret the confidence intervals—those that do not overlap zero are 'significant' (although we caution against doing so without considering the effect sizes of interest and broader context of the work (see also [44])).

2.5.2. SVM analyses

Performance of all SVM classifiers was assessed using receiver operating characteristic analysis to extract area under the curve (AUC) values. AUC is superior to classification accuracy because it characterizes the trade-off between a model's false positive rate and true positive rate at all possible decision thresholds. Accuracy, on the other hand, only characterizes performance at one decision threshold [45]. As such, AUC is a more sensitive metric for differentiating the performance of classifiers. We analyzed the performance of within- and cross-task models separately, as the test sets in both settings were of different size and different content. To compare classifiers, we used mixed-effects models, with the SVM classifier for the n-back task serving as baseline. Included in the reported models are main effects of task and classifier, and a random intercept by participant. We excluded the random slope of classifier because the statistical model failed to converge, likely indicating overfitting [46].

One limitation to our classification approach and most cognitive load classifiers in the literature is that the models treat load as binary (either being categorically high or low). However, according to subjective reports, participants experience multiple different levels of load within high- and low-load

¹For more on this, see <http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#inference-and-confidence-intervals>

conditions [2, 29, 30]. Therefore, we sought to test whether our models were sensitive to these subtle differences in subjective reports of cognitive load within high- and low-load conditions. In addition to outputting binary class labels for an arbitrary decision value of 0.5, the sliding-window SVMs also output decision values for each time point. These decision values represent the distance of the data sample from the surface the model has learned to separate the two classes. To determine whether these decision values correlated with subjective reports of mental effort, we calculated the average decision value during the previous block of trials and used it to predict the participant's subjective report of mental effort for that block of trials when controlling for the effect of our load manipulation as a covariate. Prior to the analysis, we standardized the decision values by subtracting their mean and dividing by their standard deviation across all test data. This addressed the issue that different models trained on different tasks or participants may not be well calibrated to each other. Our normalization approach allows us to interpret standardized effect sizes in this analysis.

Finally, to investigate the features that these classifiers used to discriminate periods of high- versus low-load, we carried out a set of additional analyses. First, we plotted both the time-course of standardized decision values during test blocks of high- and low-load trials to understand *when* classifiers predicted high versus low load within each type of task. Next, we trained SVMs on isolated canonical EEG frequency bands to determine *which frequencies* led to most reliable performance. Training and testing of these classifiers were done using identical procedures as described above, except these models only received EEG features from delta/theta (1–7 Hz), alpha (8–12 Hz), beta (13–30 Hz), or gamma (31–45 Hz) band frequencies. Finally, we analyzed the spatial topography of EEG activations during time windows of test data where the classifier was most confident participants were under high- or low-load to determine *where* on the scalp relevant EEG features were located. To do this we found the top 10% of high- and low-load decision values during each task period and averaged the power spectral density of the EEG over the corresponding time windows. We then averaged the activation over the relevant canonical frequency bands described above and over all participants to compare EEG activations during high- and low-load across the group.

3. Results

3.1. Higher reported mental effort, longer response times, and lower response accuracies in high-load conditions

For each of the n-back, mental arithmetic, and MOT tasks, participants were asked to report the amount

of mental effort required for each block of trials using the NASA-TLX mental effort question (table 1) [29]. A summary of the statistical analysis of the TLX scores is in table 2. The baseline TLX score for n-back was about 9.5, with mental arithmetic and MOT both scoring about two points higher. The effect of the load manipulation (i.e. the difference in score between high and low conditions) for the n-back task was about 7.6 points, with the effect larger by about 1.8 points for mental arithmetic and smaller by about 1 point for MOT. There was also a small practice effect of about 0.1 points reduction in TLX score per block. All confidence intervals exclude zero, but we suspect that the observed practice effect is not large enough to be important in practice.

In addition to TLX scores, response times and accuracies were also collected during each task (table 1). MOT response times were excluded from the analysis because participants made between 2 and 6 responses per trial, making them not directly comparable to the other tasks. The summary of the statistical analysis of the response times is in table 2. The effect of the load manipulation on n-back response times was approximately a 460 ms slowdown, while the effect of the load manipulation on mental arithmetic was about 510 ms larger. There was also a small practice effect, where participants sped up by about 10 ms per block, and a smaller fatigue effect in that participants slowed down by about 2 ms per trial within a block. As above, all confidence intervals exclude zero, but the effect of trial and block effects are likely too small to be practically relevant.

Participant performance across tasks and load conditions is presented in table 1. The summary of the statistical analysis of accuracy is in table 2. Given that correct/error choices are binary rather than normally distributed observations, generalized linear mixed models with a binomial observation likelihood and logit link function were used for this analysis. While logits are not intuitively interpretable, their relative sizes still provide useful information: Participant performance on mental arithmetic and MOT was lower than n-back (with comparable effect sizes of about -1.4 and -1.6 , respectively). The load manipulation reduced performance on the n-back task as expected (effect size of about -1.6 , which is comparable to the switch from n-back to mental arithmetic). The effect of load on MOT was comparable to the effect of load on n-back (-0.18 relative to n-back), whereas the effect of load on mental arithmetic was larger (at -1.3 relative to n-back). There was no meaningful effect of block on accuracy, as the confidence intervals included zero.

In summary, as expected, our manipulation of cognitive load was successful in evoking higher subjective reports of mental effort, longer response times, and lower task performance.

Table 1. TLX, response time, and performance across tasks and load conditions.

Task	TLX score		Response time		Accuracy	
	High load	Low load	High load	Low load	High load	Low load
N-back	12.77 (3.86)	5.16 (3.71)	1133.46 (496.05)	666.79 (248.03)	.94 (.064)	.99 (.041)
Mental Arithmetic	15.64 (3.50)	6.28 (3.64)	2034.38 (1101.99)	1053.14 (541.40)	.68 (.14)	.97 (.067)
MOT	14.30 (3.80)	7.65 (4.49)			.75 (.089)	.95 (.039)

Average subjective reports of mental effort (NASA-TLX), response time, and performance by load condition and task. Listed in parentheses are standard deviations corrected for within-participant variability [53]. Corrected standard deviations for accuracy could not be computed as the mean accuracy captures both the central tendency and variability of a binary outcome. Therefore, we report the uncorrected cross-participant standard deviations for accuracy. Response time is not reported for MOT because this task required between two and six responses and therefore are not comparable to the other two tasks.

Table 2. Effect of task and cognitive load on participant TLX, response time, and performance.

Variable	Est.	CI (lower)	CI (upper)
TLX			
Baseline (N-back)	9.54	8.28	10.76
Mental Arithmetic (difference from N-back)	2.00	1.53	2.49
MOT (difference from N-back)	2.01	1.52	2.48
Effect of load: N-back	7.59	6.33	8.78
Effect of load: Mental Arithmetic (difference from N-back)	1.75	0.77	2.77
Effect of load: MOT (difference from N-back)	-0.96	-2.04	0.12
Block	-0.11	-0.18	-0.03
Response time			
Baseline (N-back)	917.52	840.92	990.72
Mental Arithmetic (difference from N-back)	675.98	655.35	696.29
Effect of load: N-back	465.68	383.13	554.39
Effect of load: Mental Arithmetic (difference from N-back)	513.53	470.96	550.92
Block	-9.89	-12.14	-7.75
Trial (within block)	1.94	1.34	2.55
Response accuracy			
Baseline (N-back)	3.57	3.37	3.78
Mental Arithmetic (difference from N-back)	-1.42	-1.66	-1.17
MOT (difference from N-back)	-1.60	-1.76	-1.42
Effect of load: N-back	-1.62	-1.91	-1.35
Effect of load: Mental Arithmetic (difference from N-back)	-1.30	-1.85	-0.84
Effect of load: MOT (difference from N-back)	-0.18	-0.54	0.12
Block	0.01	0.00	0.03

Separate mixed-effects models were fit to investigate the effect of cognitive load manipulation and task on subjective reports of mental effort (TLX), response time, and response accuracy. For the TLX model, effect sizes are in units of TLX score, with 95% confidence interval computed by parametric bootstrap. This analysis demonstrates the success of our manipulation, namely that participants reported higher load (via TLX score) in our high-load relative to low-load condition. For the response time model, effect sizes are in milliseconds, with 95% confidence interval computed by parametric bootstrap. As mentioned in the text, we did not include MOT in this model as there are between two and six responses in each trial, and the RTs in that task are generally not comparable to the other tasks. For the response accuracy model, effect sizes are presented in logits (log-odds), with 95% confidence interval computed by parametric bootstrap. These models confirm the success of our manipulation, namely that the high load condition increased subjective reports of mental effort, increased response times and decreased response accuracy across the tasks.

3.2. SVMs reliably predict high versus low cognitive load within and across tasks

To determine whether classifiers could continuously index differences in cognitive load across time, SVMs with radial basis function (RBF) kernels were trained to predict whether sliding-windows of data belonged to either a high- or low-load trial. The input to the classifiers is the power spectral density from all 64 EEG channels from 1–45 Hz. The output is a decision value indicating whether the time window belonged to either a high- or low-load trial (see section 2).

3.2.1. Within-task performance

First, we trained models on data from a subset of blocks of one task (mental arithmetic, n-back, or MOT) and tested the models on data from different blocks of the same task in a five-fold cross-validation procedure. SVMs were able to reliably predict whether 3 s sliding-windows of data belonged to high- or low-load trials for all three tasks across participants (table 3 and figure 3). SVM classifiers trained on 3 s windows of activity were used for the majority of the analyses since they performed comparably to classifiers trained

on longer windows (figure S1 (available online at stacks.iop.org/JNE/17/056016/mmedia)), while maintaining increased interpretability about which time points during each task lead to high cognitive load predictions.

Table 3 gives detailed results of classification performance for each task. The mean AUC of SVMs applied to the n-back task was .73. AUC for SVMs applied to mental arithmetic and MOT tasks were .04 and .09 lower, respectively. The other tested classifiers, kNN with 9 neighbors and LDA, performed about .07 and .08 lower in AUC, respectively. In sum, the RBF SVM was consistently the best-performing of the tested classifiers for within-task classification, with a notable difference for practical purposes of about 10%. To facilitate comparisons with existing literature, classification accuracy is also presented in table 3 despite its decreased sensitivity in detecting differences between classifiers (see section 2). Results for kNN with different choices of k are presented in figure S2. There were marginal improvements in kNN performance when using $k = 9$ relative to $k = 5$ or $k = 3$ across all tasks.

Together, these results demonstrate that sliding-window SVM models were able to reliably predict cognitive load within all three of the experimental tasks and outperformed other commonly used classification algorithms.

3.2.2. Cross-task performance

Next, sliding-window SVMs were tested to determine if they could reliably discriminate between periods of high- and low-load during tasks they had not been trained on. Ultimately, the ability to classify across tasks is important for practical applications of models used to monitor cognitive load in the real world. Assuming successful cross-task performance, these models can also be used to identify EEG features that reflect general cognitive load rather than task-specific confounds. To this end, models were trained on data from two of the three tasks and tested on data from the held-out task (see section 2).

Cross-task SVM models performed significantly above chance for each of the three held-out tasks (table 3 and figure 3). Differences in AUC for cross-task performance were similar to those for within-task. Models tested on n-back had an AUC of .75, similar to those tested on mental arithmetic, while models tested on MOT performed .1 lower in AUC. The other classifiers, LDA and kNN, performed considerably worse than SVMs, with loss of about .1 in AUC. We caution against making strong inferences based on the similarity of across and within-task SVM performance given the differences in training and test set sizes but highlight that SVMs were able to successfully decode cognitive load across tasks. Furthermore, these classifiers outperformed the other tested classifiers conventionally used for cognitive load prediction on EEG data.

3.3. Model outputs capture differences in the time-course of cognitive load across tasks

Given that the cross-task SVM models were able to reliably discriminate periods of high- and low-load, we sought to determine whether their outputs captured continuous fluctuations in load that occur within individual trials. By looking at the time-course of the models' output for the test blocks, which is a standardized decision value over time, the confidence of the model can be visualized through time. The outputs of the cross-task models are illustrated for one representative high- and low-load test block for each task averaged over participants in figure 4. The average output during all cross-task test blocks are illustrated in figure S3.

Within each task, high-load blocks were consistently evaluated as having higher decision values than low-load blocks, which is expected given that the models demonstrated above chance AUC for high-versus low-load classification. However, the time-course of cognitive load fluctuations during trials of each task were distinct. Continuous outputs during high-load mental arithmetic blocks (figure 4) demonstrated sharp increases of load after the second integer of each sum was presented, which is when participants would begin adding the first and second presented integers. It then sharply fell off after the proposed answer was presented. During the n-back task, classifier outputs were sustained, rising to a high or low level at the beginning of each block and remaining constant throughout the block. Finally, the MOT tasks elicited sustained increases in load during the entire tracking phase and persisted until after responses were indicated at the end of each trial.

In summary, the continuous outputs of these cross-task sliding-window SVMs suggest differences in the time-courses of cognitive load changes within trials of each task.

3.4. Model outputs capture subtle differences in subjective reports of mental effort

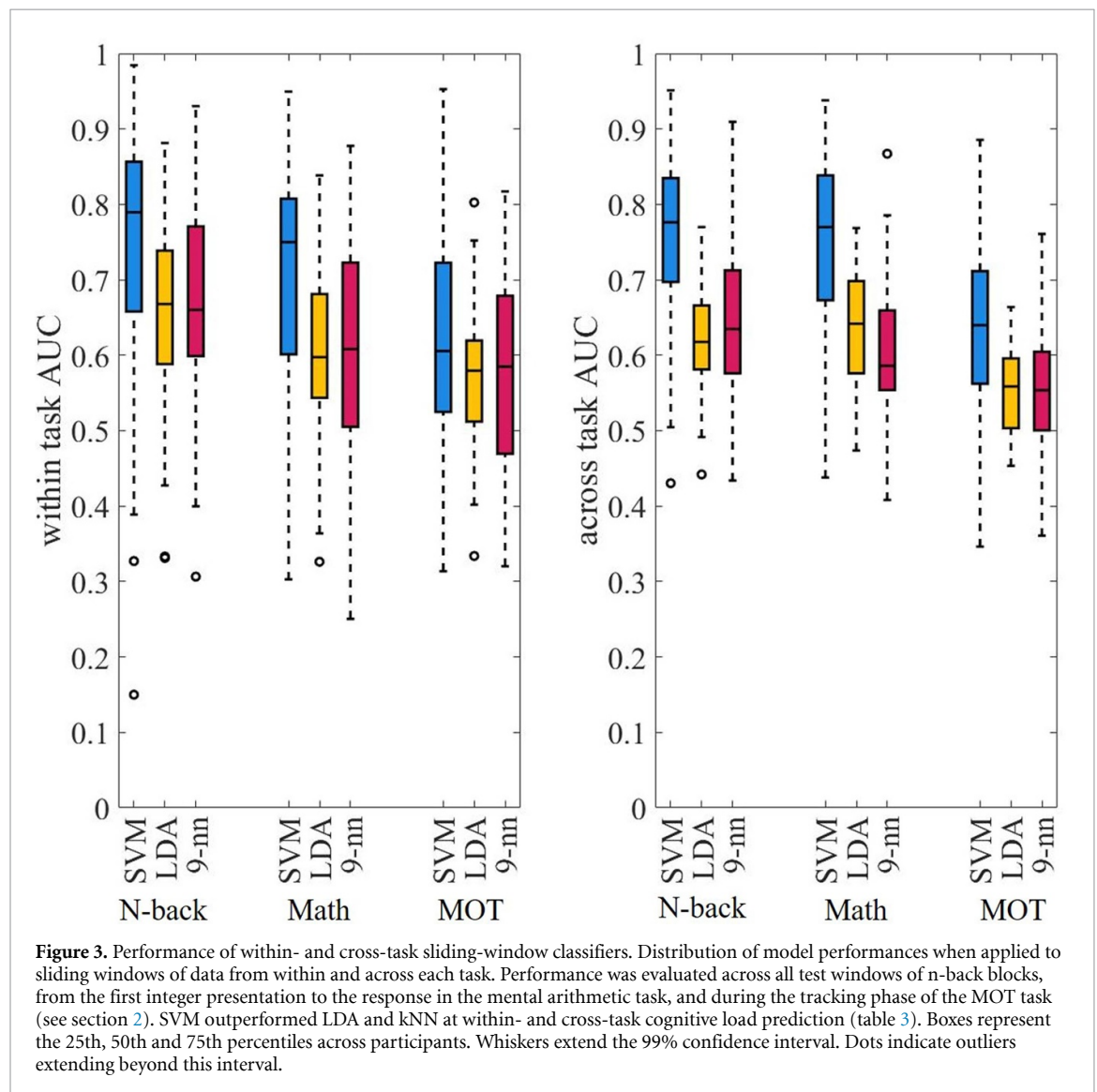
To validate this continuous index of cognitive load, the decision values of the model were tested to determine if they were sensitive to differences in reported mental effort within high- and low-load conditions. Specifically, the average decision value of the cross-task classifier during each block of the test task was computed. These average decision values represent how confident the classifier was that the participant was under high- or low-load during that block. This average decision value was added as a covariate of TLX score to the model reported in table 2, as well as its interaction with the load manipulation. The results of this model are summarized in table 4 and the correlation between TLX score and model outputs is illustrated in figure 5.

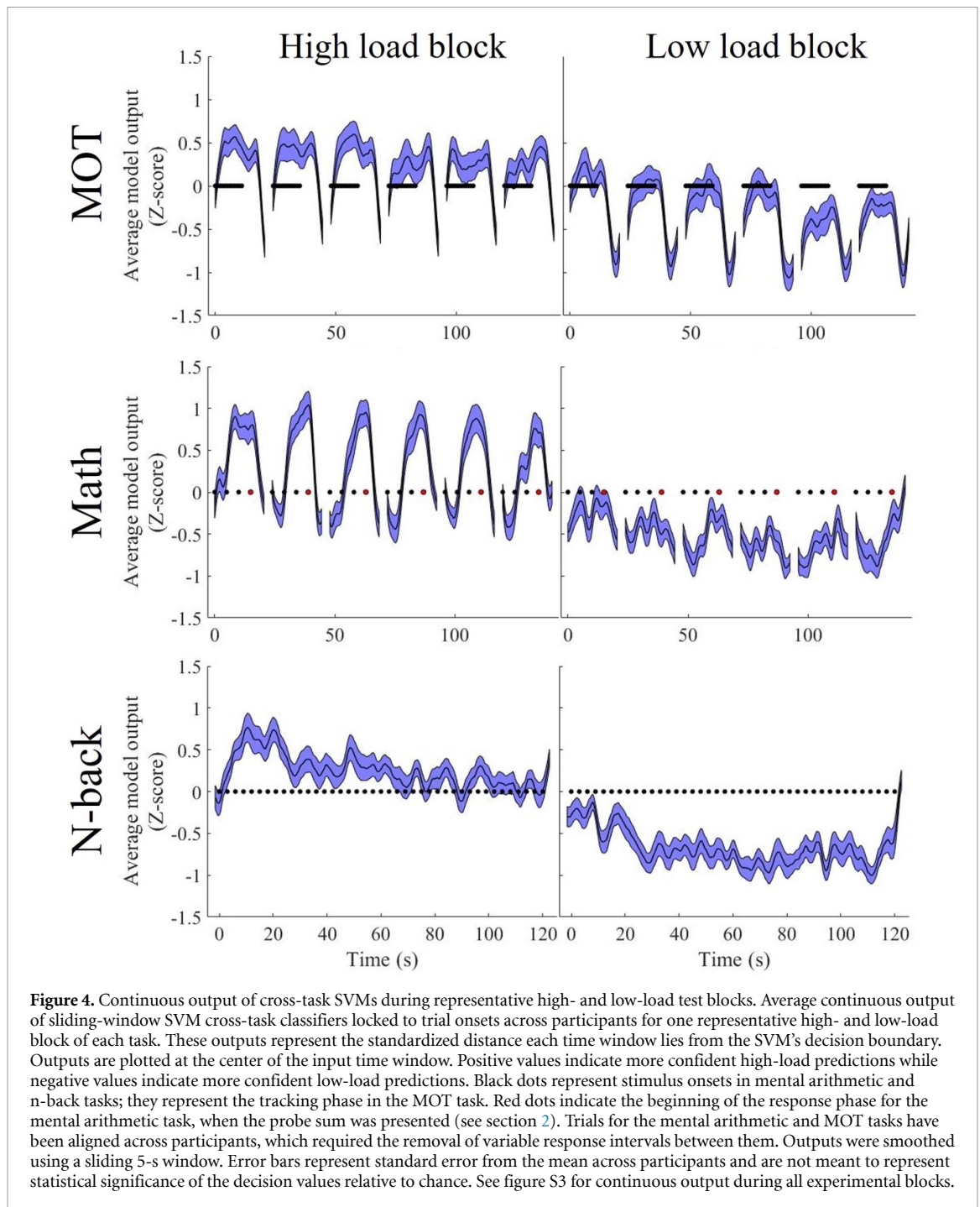
Even in the presence of a main effect of the load manipulation on TLX score, there was an effect of

Table 3. Analysis of within- and cross-task classification performance.

Variable	AUC			Accuracy		
	Est.	CI (lower)	CI (upper)	Est.	CI (lower)	CI (upper)
Within-Task Model						
Baseline: N-back, SVM	.73	.69	.77	.67	.64	.70
Mental arithmetic (difference from N-back)	-.04	-.07	-.02	-.03	-.05	-.02
MOT (difference from N-back)	-.09	-.12	-.07	-.08	-.10	-.06
KNN classifier (relative to SVM)	-.07	-.09	-.05	-.07	-.08	-.05
LDA classifier (relative to SVM)	-.08	-.10	-.06	-.03	-.05	-.01
Cross-Task Model						
Baseline: N-back, SVM	.75	.72	.78	.67	.64	.69
Mental Arithmetic (difference from N-back)	-.01	-.03	.01	.00	-.01	.02
MOT (difference from N-back)	-.10	-.12	-.08	-.08	-.09	-.06
KNN classifier (relative to SVM)	-.11	-.13	-.09	-.08	-.10	-.06
LDA classifier (relative to SVM)	-.11	-.13	-.09	-.04	-.05	-.02

Separate mixed effects models were computed to compare performance for within- and cross-task classification. We caution against making direct comparisons for across versus within-task classifiers due to differences in training set sizes and cross-validation procedures. Effect sizes are in units of AUC (left) and accuracy (right), with 95% confidence interval computed by parametric bootstrap. Not shown is a random intercept by participant. This analysis demonstrates that best of the classifier we tried was the RBF SVM, with the best performance on the n-back task for both within- and cross-task classification. We show a minor (likely negligible) performance drop on the mental arithmetic task, and a greater drop from using one of the other classifiers (kNN with $k = 9$ or LDA) and for decoding load on the MOT task both within- and cross-task.





about 0.6 points of the standardized SVM decision value on reported TLX score. That is, each standard deviation of the SVM's decision value corresponded to about a 0.6-point difference in TLX score, even when controlling for the effect of the load manipulation (which was almost an order of magnitude higher at about 7 points). The effect size of the interaction between the SVM decision value and the load manipulation was nearly zero, indicating that the predictive ability of the SVM outputs did not depend on high- versus low-load condition. This analysis demonstrates that cross-task SVM models were able to capture subtle differences in subjectively reported mental effort within high and low load conditions,

even without explicit training on these subjective reports.

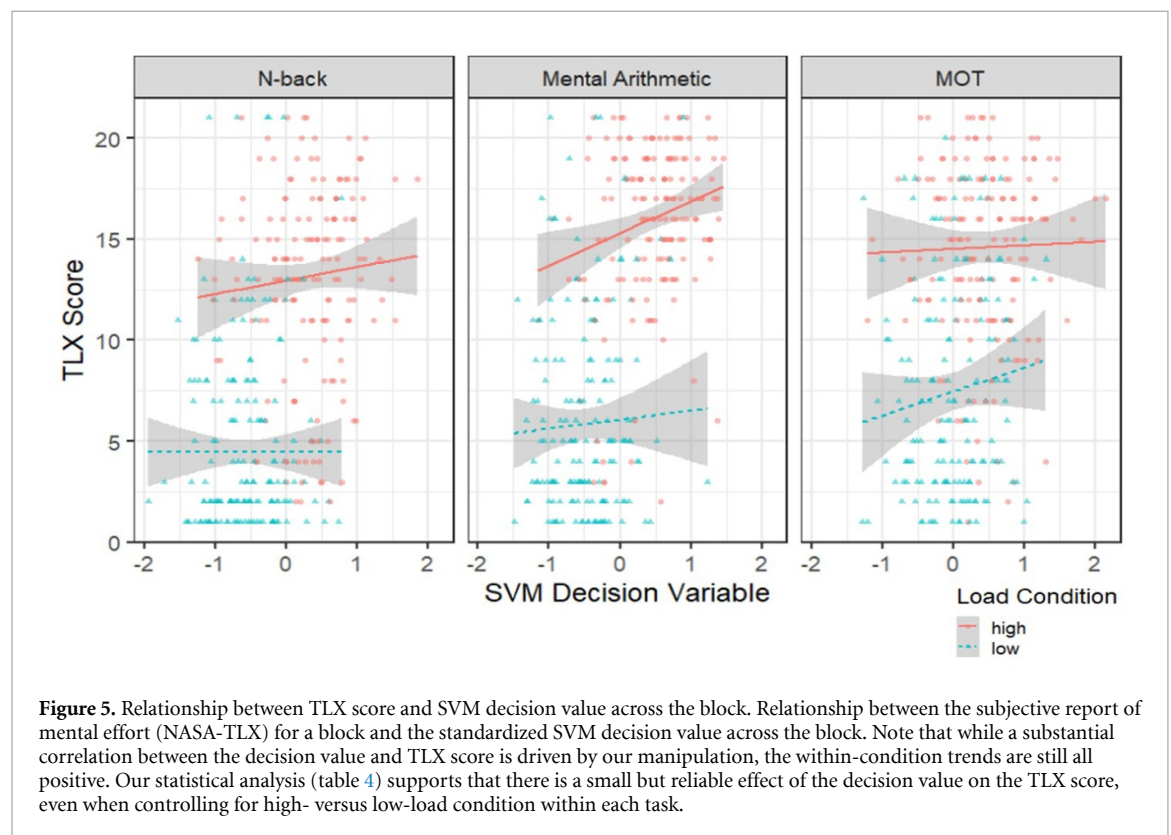
3.5. Task-general and task-specific EEG correlates of cognitive load

To better understand the neural features that drove performance of the cross-task SVM models, we compared the performance of classifiers trained only on features from individual canonical EEG frequency bands (delta/theta: 1–7 Hz, alpha: 8–12 Hz, beta: 13–20 Hz, and gamma: 31–45 Hz). Models trained on only alpha or beta band features performed comparably to those trained on features from all frequency bands for both within- and cross-task load prediction

Table 4. Analysis of TLX score across tasks and conditions with SVM decision value added.

Variable	Est.	CI (lower)	CI (upper)
Baseline score: N-back	9.54	8.22	10.82
Mental Arithmetic (difference from N-back)	1.93	1.38	2.38
MOT (difference from N-back)	1.77	1.23	2.31
Effect of load: N-back	7.01	5.72	8.37
Effect of load: Mental Arithmetic (difference from N-back)	1.72	0.61	2.73
Effect of load: MOT (difference from N-back)	-0.61	-1.74	0.58
SVM decision value	0.64	0.23	1.08
SVM decision value-load interaction	-0.15	-0.96	0.75
Block	-0.08	-0.17	0.00

TLX mixed-effect model presented in table 2 with the added covariate of SVM decision value. Effect sizes are in units of TLX score, with 95% confidence interval computed by parametric bootstrap. Not shown are random effects of participant (random intercept, plus random slope of load manipulation). There is a small effect of the SVM decision value in predicting TLX score even when controlling for our load manipulation, i.e. the classifier's output is predictive of load continuously within each load condition.



(table 5, figure 6). Models trained only on gamma activity, however, suffered from a drop of -0.14 and -0.11 in AUC relative to the full model at both within and across task load prediction, respectively. Finally, models trained on only theta/delta band activity suffered from a drop of -0.12 AUC relative to the full cross-task models compared to a drop of -0.04 AUC compared to the full within-task models. These results suggest that alpha and beta frequencies contained the most relevant information for within- and cross-task cognitive load classification, whereas theta activity was more reflective of cognitive load within-task.

Next, to provide insight into the spatial EEG features that lead to high and low cognitive load model predictions, the average EEG activation that produced the top and bottom 10% of the full cross-task models' decision values were plotted. These

time-points correspond to the those that the classifiers were most confident belong to high- or low-cognitive load conditions. The trial periods that were most represented by these high- and low-load predictions can be seen in figure S4. The EEG activations within frequency bands with fair within or cross-task classification performance (theta, alpha, and beta) were averaged across these time points and across participants to illustrate the spatial patterns of activity that lead to high- and low-load predictions within each task (figure 7).

These EEG activation maps revealed differences in the patterns of activity elicited by the three tasks. For example, average EEG power was suppressed in theta and alpha frequency bands in MOT relative to n-back and mental arithmetic during both high- and low-load conditions. However, across the three tasks,

Table 5. Comparing the performance of different frequency bands for within- and cross-task load classification.

Variable	AUC			Accuracy		
	Est.	CI (lower)	CI (upper)	Est.	CI (lower)	CI (upper)
Within-Task Model						
Baseline: N-back	.73	.69	.77	.67	.64	.70
Mental Arithmetic (difference from N-back)	-.05	-.07	-.03	-.03	-.05	-.02
MOT (difference from N-back)	-.09	-.11	-.07	-.07	-.08	-.06
Theta band (relative to all bands)	-.04	-.06	-.02	-.03	-.05	-.01
Alpha band (relative to all bands)	.03	.00	.05	.02	.00	.04
Beta band (relative to all bands)	-.01	-.04	.01	-.01	-.03	.01
Gamma band (relative to all bands)	-.14	-.17	-.12	-.11	-.13	-.09
Cross-Task Model						
Baseline: N-back	.75	.71	.77	.66	.64	.68
Mental Arithmetic (difference from N-back)	.00	-.01	.02	.02	.00	.03
MOT (difference from N-back)	-.09	-.11	-.07	-.07	-.08	-.05
Theta band (relative to all bands)	-.12	-.14	-.09	-.07	-.09	-.06
Alpha band (relative to all bands)	-.03	-.05	-.01	-.02	-.04	.00
Beta band (relative to all bands)	-.02	-.04	.00	-.01	-.03	.00
Gamma band (relative to all bands)	-.11	-.14	-.09	-.07	-.09	-.06

We fit separate mixed-effects models to investigate the relative performance of different frequency bands for within- and cross-task classification (delta/theta: 1–7 Hz, alpha: 8–12 Hz, beta: 13–30 Hz, and gamma: 31–45 Hz). We caution against making direct comparisons for across versus within-task classifiers due to differences in training set sizes and cross-validation procedures. Effect sizes are in units of AUC (left) and accuracy (right), with 95% confidence interval computed by parametric bootstrap. Not shown is a random intercept by participant. For within-task classification, models trained on only alpha or beta activity perform comparably to models trained on all frequency bands. There's a small decrease in performance when using only theta activity but a large drop when using only gamma activity. For cross-task classification, models trained only on alpha and beta activity also perform comparably to models trained on all bands. However, models trained only on gamma and theta activity show a large drop in performance.

periods of high load were marked by relatively lower activation in alpha and beta power relative to periods of low load. Changes in theta associated with high cognitive load were not consistent across tasks. For n-back and mental arithmetic, periods of high load were associated with increased frontal theta activity, whereas there was an overall decrease in theta during high load MOT trials. There were no clear differences between the trial periods that contributed to high-versus low-load activation maps within each task (figure S4). Therefore, it is unlikely that differences in high- versus low-load activation maps were reflective of different trial phases (i.e. stimulus onset versus offset).

Together these results support previous evidence that alpha and beta band activity may be most relevant for cross-task load prediction [25].

4. Discussion

Cognitive load continuously fluctuates over time and has been shown to contribute to errors when excessively high [3–5]. Therefore, models that can reliably identify periods of high load and use that information to mitigate potential errors have the potential to greatly improve human performance. However, to be maximally useful, these models must be able to reliably index cognitive load continuously over time in a variety of different environmental and task contexts. In other words, it is critical that these models generalize across changes in cognitive, perceptual,

and motor features that are not related to cognitive load.

The sliding-window SVMs presented here produced several key results. First, they were able to generalize to tasks that they had not been trained on (figure 3). Second, these cross-task models produced continuous estimates of cognitive load over time, which provided insight into the time-course of cognitive load changes within individual trials of commonly used laboratory tasks (figure 4). Third, these continuous model outputs captured subtle differences in subjective reports of mental effort within high- and low-load conditions (figure 5). Finally, sliding-window SVMs trained on isolated EEG frequency bands demonstrated which frequency bands reflect task-general and task-specific correlates of cognitive load (figures 6 and 7). Thus, these algorithms show promise for detecting subtle changes in cognitive load in real-time across a variety of contexts and provide insight into the task-general EEG correlates that reflect changes in cognitive load. Below we provide potential interpretations for (1) differences in the time-course of cognitive load fluctuations within trials of different tasks and (2) the underlying task-general and task-specific EEG activations that were learned by the models.

4.1. Differences in the time-course of cognitive load fluctuations across tasks

A key finding from these analyses was that different tasks evoked changes in cognitive load with different time-courses. For example, the n-back task

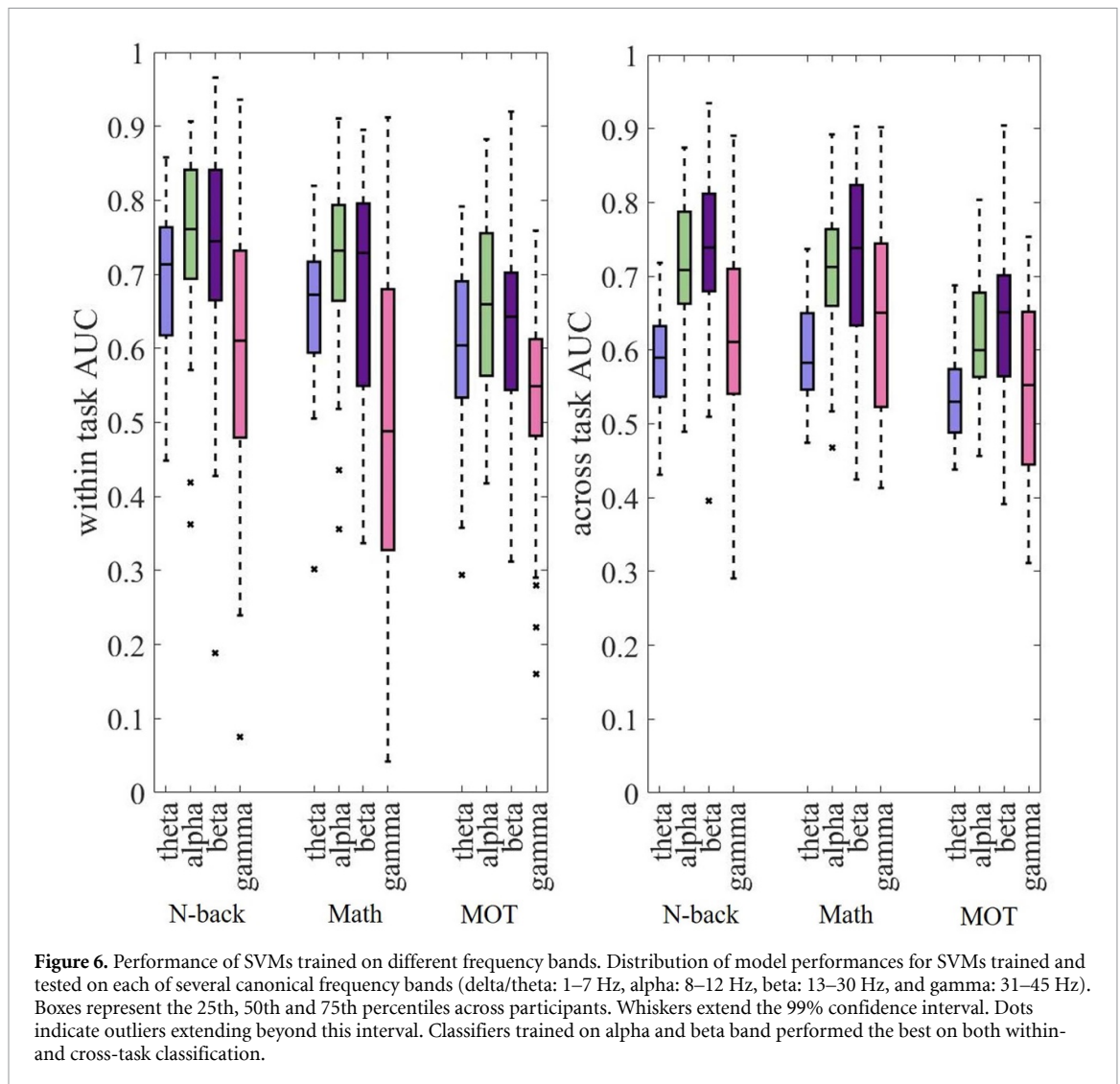


Figure 6. Performance of SVMs trained on different frequency bands. Distribution of model performances for SVMs trained and tested on each of several canonical frequency bands (delta/theta: 1–7 Hz, alpha: 8–12 Hz, beta: 13–30 Hz, and gamma: 31–45 Hz). Boxes represent the 25th, 50th and 75th percentiles across participants. Whiskers extend the 99% confidence interval. Dots indicate outliers extending beyond this interval. Classifiers trained on alpha and beta band performed the best on both within- and cross-task classification.

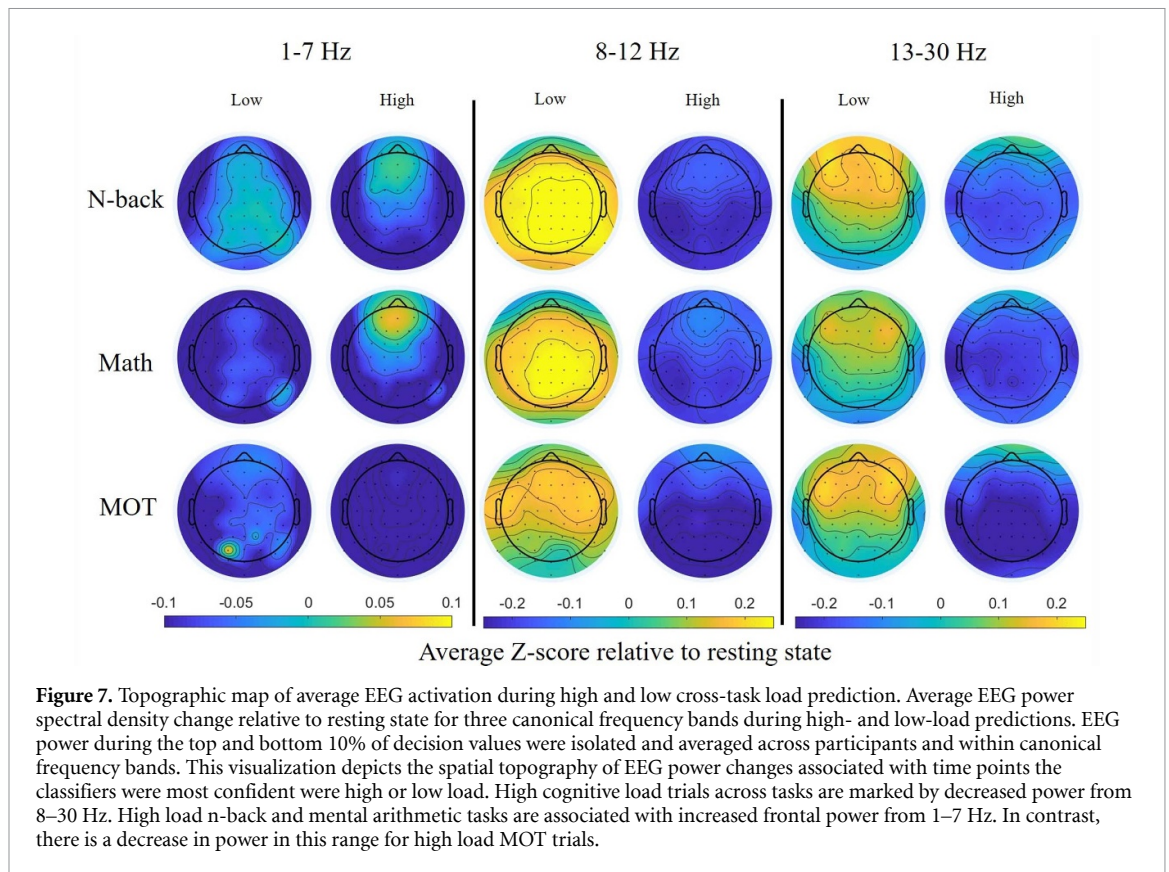
requires sustained maintenance and updating of working memory throughout a high-load block, whereas mental arithmetic allows for dips in cognitive load between trials because it is not necessary to carry information over to the next trial. Sliding-window SVMs were able to detect these predicted fluctuations in cognitive load during tasks that they were not trained on (figure 4).

For example, mental arithmetic elicited increases in cognitive load that lasted from the second integer through the final test of the proposed sum. These increases may reflect increased cognitive load associated with carrying numbers during the high-load task compared to the low-load task where this transient increase in load was not as great [47]. For MOT, cognitive load was sustained throughout the duration of the tracking period for both high- and low-load conditions and dropped between trials. This finding suggests that splitting exogenous attention across the targets during the tracking phase and maintaining that attention during the response phase increased load during high- and low-load versions of this task, although to a greater degree in the

high-load condition. On the other hand, during n-back blocks, there was a consistent and sustained increase or decrease in load, which began at the start of each block and lasted throughout. This is consistent with the idea that working memory is continuously updated through the block of trials as one must maintain the memory of the letter from prior trials [22].

4.2. Task-general EEG correlates of cognitive load

Several underlying cognitive states, like working memory and attention, have the potential to affect cognitive load differently across tasks [14, 15]. The current study helps illustrate the task-general EEG correlates that reflect these changes in cognitive load. Further, this work demonstrates that these task-general correlates of cognitive load manifest even across tasks that differentially affect participant performance and reaction times. For example, the high-load n-back task was rated by participants as requiring significantly higher mental effort compared to the low-load n-back task, even though response accuracy for the high-load n-back task was much



higher than the response accuracy for the high-load mental arithmetic and MOT tasks. Despite these differences in behavioral performance across the tasks, sliding window SVMs were still able to reliably predict the cognitive load of participants.

Models trained solely on alpha (8–12 Hz) or beta (13–30 Hz) band features performed better at cross-task cognitive load prediction than those trained on only delta/theta (1–7 Hz) or gamma (31–45 Hz) band features (figure 6). Furthermore, when examining the EEG activation patterns during windows with highest and lowest model decision values, there was decreased parietal alpha and frontal beta activity for high-load windows compared to low-load windows across all tasks. This result is consistent with several reports that changes in alpha [12, 14, 20, 48, 49] and beta [14, 28] activity is reflective of changes in cognitive load.

Interestingly, although changes in theta band activity have often been reported during experimental manipulations of cognitive load [9, 12, 14, 20, 50, 51], models trained on EEG features in the range of 1–7 Hz performed worse than any other frequency band at cross-task load classification. However, within-task models trained only on 1–7 Hz activity performed moderately well. This difference between within-versus cross-task performance suggests that the spatial topography or direction of theta band changes associated with high load may vary from task to task. This is supported by the observed EEG activations during high- and low-load predictions across

the tasks (figure 7). In mental arithmetic and n-back tasks, increases in frontal theta occurred during high load periods, whereas decreases in theta occurred during high load periods of MOT.

This may also partially explain why within- and cross-task models performed consistently worse at classifying MOT trials relative to the other two tasks. During the MOT task, participants were asked to modulate their exogenous attention to follow the movement of either two or six moving targets [32], whereas n-back and mental arithmetic tasks targeted endogenously maintained information [22, 31]. These two processes—endogenous and exogenous attention—have been associated with distinct EEG signatures [50, 51]. Given that both within- and cross-task SVM models performed significantly worse when tested on MOT data compared to n-back or mental arithmetic, it may be the case that the models were less sensitive to the EEG features associated with changes in exogenous attention relative to endogenous attention.

Another potential explanation for decreased classification performance on MOT is that participants were fatigued during the MOT task, which always occurred last. However, continuous model outputs, which show clear increases in cognitive load during the tracking phase of MOT and sharp decrease after the conclusion of each block (figure 4), suggest that participants were engaging in the task. In summary, the models presented here suggest that task-general EEG signatures of cognitive load are reflected in alpha

and beta frequencies, whereas underlying changes in attention that varied across the tasks studied primarily manifested in lower frequency activity.

4.3. Future directions

In this work we were able to design cognitive load prediction algorithms that could generalize across of variety of perceptually and cognitively distinct tasks. Future work might enhance these prediction algorithms in several ways. For example, future efforts to model cognitive load might focus on detecting small and sudden changes in cognitive load rather than changes in load that persist over several seconds of an experimental trial (3 s in this study). It remains to be determined whether current predictive models of cognitive load are sensitive to rapid fluctuations because most studies use seconds of data to produce a single output. Also, it is unclear how variable task-general correlates of cognitive load are across participants and whether this variation corresponds to differences in cognitive ability [8]. Understanding cross-participants differences could lead to cognitive load prediction algorithms that do not require as much data to tailor to individuals, while increasing the amount of data that can be pooled across participants. Finally, future work might enhance cognitive load prediction by leveraging history to make better inferences of load. The models presented here are trained on each sliding-window of data independently, and as such they are unable to take advantage of history when making their predictions. In this regard, convolutional neural network models used to learn spatiotemporal representations, or other sequential statistical models, may prove advantageous because they may be able to learn features corresponding to both sustained and quickly fluctuating changes in load [18, 26, 52]. However, to date these models have only been trained to make trial-level predictions, and scaling them up to be temporally continuous remains difficult due to the computational cost of tuning the parameters of these large neural networks [40].

5. Conclusion

Designing models that can robustly capture task-general changes in cognitive load in the real world remains an important challenge. In the real world, people complete a variety of different ‘tasks’ that induce changes in cognitive load. These tasks can evolve on different timescales and evoke large changes in EEG features associated with the task’s cognitive, motor, or perceptual features.

The sliding-window SVM models presented here allow for continuous monitoring of cognitive load over time across a variety of laboratory tasks, which marks an important step toward continuous monitoring of cognitive load in the real world. These models perform well at both within-task cognitive load prediction and cross-task generalization, and

model outputs are sensitive to differences in subjective reports of mental effort. Therefore, these types of models are promising for continuous cognitive load detection in the real world, which can be used to prevent potentially costly errors and improve human performance.

Acknowledgments

The authors would like to thank Anjali Misra and Hannah Postings for assisting with data collection, Qiong Zhang for sharing pre-processing scripts, and Brian Hecox for software support. Additionally, the authors acknowledge Emily Mugler, Amir Memar, Ying Yang, Sho Nakagome, and Xueqing Liu for their valuable feedback on the analyses and manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

ORCID iD

Matthew J Boring  <https://orcid.org/0000-0002-6099-4815>

References

- [1] Leppink J, Paas F, van Gog T, van der Vleuten C P M and van Merriënboer J J G 2014 Effects of pairs of problems and examples on task performance and different types of cognitive load *Learn. Instr.* **30** 32–42
- [2] Paas F G 1992 Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* **84** 429–34
- [3] Horrey W J and Wickens C D 2006 Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Hum. Factors* **48** 196–205
- [4] Parasuraman R 2011 Neuroergonomics: brain, cognition, and performance at work *Curr. Dir. Psychol. Sci.* **20** 181–6
- [5] Shappell S, Detwiler C, Holcomb K, Hackworth C, Boquet A and Wiegmann D A 2017 Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system *Human Error in Aviation* (Oxford: Routledge) pp 73–88
- [6] Wilson G F and Russell C A 2003 Real-time assessment of mental workload using psychophysiological measures and artificial neural networks *Hum. Factors J. Hum. Factors Ergon. Soc.* **45** 635–44
- [7] Gerjets P, Walter C, Rosenstiel W, Bogdan M and Zander T O 2014 Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach *Front. Neurosci.* **8** 385
- [8] Wang Z, Hope R M, Wang Z, Ji Q and Gray W D 2012 Cross-subject workload classification with a hierarchical Bayes model *Neuroimage* **59** 64–69
- [9] Puma S, Matton N, Paubel P-V, Raufaste É and El-Yagoubi R 2018 Using theta and alpha band power to assess cognitive workload in multitasking environments *Int. J. Psychophysiol.* **123** 111–20
- [10] Walter C, Schmidt S, Rosenstiel W, Gerjets P and Bogdan M 2013 Using cross-task classification for classifying workload levels in complex learning tasks *Proc. - 2013 Humaine Association Conf. on Affective Computing and Intelligent Interaction, ACII 2013* pp 876–81

- [11] Gevins A, Smith M E, Leong H, McEvoy L, Whitfield S, Du R and Rush G 1998 Monitoring working memory load during computer-based tasks with EEG pattern recognition methods *Hum. Factors J. Hum. Factors Ergon. Soc.* **40** 79–91
- [12] Spüler M, Walter C, Rosenstiel W, Gerjets P, Moeller K and Klein E 2016 EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning *ZDM - Math. Educ.* **48** 267–78
- [13] Brouwer A-M, Hogervorst M A, van Erp J B F, Heffelaar T, Zimmerman P H and Oostenveld R 2012 Estimating workload using EEG spectral power and ERPs in the n-back task *J. Neural Eng.* **9** 045008
- [14] Astrand E 2018 A continuous time-resolved measure decoded from EEG oscillatory activity predicts working memory task performance *J. Neural Eng.* **15** 036021
- [15] Kardan O, Adam K C S, Mance I, Churchill N W, Vogel E K and Berman M G 2020 Distinguishing cognitive effort and working memory load using scale-invariance and alpha suppression in EEG *Neuroimage* **211** 116622
- [16] Ke Y, Qi H, He F, Liu S, Zhao X, Zhou P, Zhang L and Ming D 2014 An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task *Front. Hum. Neurosci.* **8** 703
- [17] Baldwin C L and Penaranda B N 2012 Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification *Neuroimage* **59** 48–56
- [18] Bashivan P, Rish I, Yeasin M and Codella N 2016 Learning representations from EEG with deep recurrent-convolutional neural networks *4th Int. Conf. on Learning Representations, ICLR 2016 - Conf. Track Proc.*
- [19] Christensen J C, Estep J R, Wilson G F and Russell C A 2012 The effects of day-to-day variability of physiological data on operator functional state classification *Neuroimage* **59** 57–63
- [20] Gevins A, Smith M E, McEvoy L and Yu D 1997 High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice *Cereb. Cortex* **7** 374–85
- [21] Hogervorst M A, Brouwer A-M and van Erp J B F 2014 Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload *Front. Neurosci.* **8** 322
- [22] Owen A M, McMillan K M, Laird A R and Bullmore E 2005 N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies *Hum. Brain Mapp.* **25** 46–59
- [23] Walter C, Rosenstiel W, Bogdan M, Gerjets P and Spüler M 2017 Online EEG-based workload adaptation of an arithmetic learning environment *Front. Hum. Neurosci.* **11** 286
- [24] Sternshein H, Agam Y and Sekuler R 2011 EEG correlates of attentional load during multiple object tracking *PLoS One* **6** e22660
- [25] Sweller J 1988 Cognitive load during problem solving: effects on learning *Cogn. Sci.* **12** 257–85
- [26] Zhang P, Wang X, Zhang W and Chen J 2019 Learning spatial-spectral-temporal EEG features with recurrent 3D convolutional neural networks for cross-task mental workload assessment *IEEE Trans. Neural Syst. Rehabil. Eng.* **27** 31–42
- [27] Pfurtscheller G, Brunner C, Schlögl A and Lopes da Silva F H 2006 Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks *Neuroimage* **31** 153–9
- [28] Dimitrakopoulos G N, Kakkos I, Dai Z, Lim J, deSouza J J, Bezerianos A and Sun Y 2017 Task-independent mental workload classification based upon common multiband EEG cortical connectivity *IEEE Trans. Neural Syst. Rehabil. Eng.* **25** 1940–9
- [29] Hart S G and Staveland L E 1988 Development of NASA-TLX (Task Load Index): results of empirical and theoretical research *Adv. Psychol.* **52** 139–83
- [30] Paas F G W C and Van Merriënboer J J G 1993 The efficiency of instructional conditions: an approach to combine mental effort and performance measures *Hum. Factors* **35** 737–43
- [31] Ashcraft M H and Battaglia J 1978 Cognitive arithmetic: evidence for retrieval and decision processes in mental addition *J. Exp. Psychol. Hum. Learn. Mem.* **4** 527–38
- [32] Cavanagh P and Alvarez G A 2005 Tracking multiple targets with multifocal attention *Trends Cogn. Sci.* **9** 349–54
- [33] Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Methods* **134** 9–21
- [34] Mullen T R, Kothe C A E, Chi Y M, Ojeda A, Kerth T, Makeig S, Jung T-P and Cauwenberghs G 2015 Real-time neuroimaging and cognitive monitoring using wearable dry EEG *IEEE Trans. Biomed. Eng.* **62** 2553–67
- [35] Chang C Y, Hsu S H, Pion-Tonachini L and Jung T P 2018 Evaluation of artifact subspace reconstruction for automatic EEG artifact removal *Proc. of the Annual Int. Conf. of the IEEE Eng. Med. Biol. Soc. (EMBC)* vol 2018 pp 1242–5
- [36] Cohen L 2001 The uncertainty principle for the short-time Fourier transform and wavelet transform *Wavelet Transforms and Time-Frequency Signal Analysis* (Basel: Birkhäuser) pp 217–32
- [37] Chang C-C and Lin C-J 2011 LIBSVM: A library for support vector machines *ACM Trans. Intell. Syst. Technol.* **2** 27
- [38] Vert J P, Tsuda K and Scholkopf B 2019 A primer on Kernel methods *Kernel Methods in Computational Biology* (Cambridge, MA: The MIT Press) 35–70
- [39] Chih-Wei Hsu C-J L and Chang C-C 2008 A practical guide to support vector classification *BJU Int.* **101** 1396–400
- [40] Garrett D, Peterson D A, Anderson C W and Thaut M H 2003 Comparison of linear, nonlinear, and feature selection methods for EEG signal classification *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 141–4
- [41] Carey V J and Wang Y-G 2001 Mixed-effects models in S and S-Plus *J. Am. Stat. Assoc.* **96** 1135–6
- [42] Kane M J and Engle R W 2002 The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: an individual-differences perspective *Psychon. Bull. Rev.* **9** 637–71
- [43] Chung Y, Rabe-Hesketh S, Dorie V, Gelman A and Liu J 2013 A nondegenerate penalized likelihood estimator for variance parameters in multilevel models *Psychometrika* **78** 685–709
- [44] Wasserstein R L and Lazar N A 2016 The ASA's statement on p-values: context, process, and purpose *Am. Stat.* **70** 129–33
- [45] Huang J and Ling C X 2005 Using AUC and accuracy in evaluating learning algorithms *IEEE Trans. Knowl. Data Eng.* **17** 299–310
- [46] Bates D, Kliegl R, Vasishth S and Baayen H 2015 Parsimonious mixed models arXiv:1506.04967
- [47] Geary D C and Widaman K F 1987 Individual differences in cognitive arithmetic *J. Exp. Psychol. Gen.* **116** 154–71
- [48] Glass A and Kwiatkowski A W 1970 Power spectral density changes in the EEG during mental arithmetic and eye-opening *Psychologische Forschung* **33** 85–99
- [49] Ray W J and Cole H W 1985 EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes *Science* **228** 750–2
- [50] Keller A S, Payne L and Sekuler R 2017 Characterizing the roles of alpha and theta oscillations in multisensory attention *Neuropsychologia* **99** 48–63
- [51] Missonnier P, Deiber M-P, Gold G, Millet P, Gex-Fabry Pun M, Fazio-Costa L, Giannakopoulos P and Ibáñez V 2006 Frontal theta event-related synchronization: comparison of directed attention and working memory load effects *J. Neural Transm.* **113** 1477–86
- [52] Hefron R G, Borghetti B J, Christensen J C and Schubert Kabban C M 2017 Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation *Pattern Recognit. Lett.* **94** 96–104
- [53] Morey R D 2008 Confidence intervals from normalized data: a correction to Cousineau (2005) *Tutorials in Quantitative Methods for Psychology* **4** 61–4