EEG-Based Cognitive Load Classification During Landmark-Based VR Navigation

Jiahui An
Institute of Neuroinformatics,
University of Zurich and ETH Zurich
Zurich, Switzerland

Bingjie Cheng
Department of Geography, University
of Zurich
Zurich, Switzerland

Dmitriy Rudyka
Department of Psychology, University
of Zurich
Zurich, Switzerland

Elisa Donati
Institute of Neuroinformatics,
University of Zurich and ETH Zurich
Zurich, Switzerland

Sara Fabrikant
Department of Geography and Digital
Society Initiative, University of
Zurich
Zurich, Switzerland

Abstract

Brain–computer interfaces enable real-time monitoring of cognitive load, but their effectiveness in dynamic navigation contexts is not well established. Using an existing Virtual Reality (VR) navigation dataset, we examined whether Electroencephalography (EEG) signals can classify cognitive load during map-based wayfinding and whether classification accuracy depends more on task complexity or on individual traits. EEG recordings from forty-six participants navigating routes with 3, 5, or 7 map landmarks were analyzed with a nested cross-validation framework across multiple machine learning models. Classification achieved mean accuracies up to 90.8% for binary contrasts (3 vs. 7 landmarks) and 78.7% for the three-class problem, both well above chance. Demographic and cognitive variables (age, gender, spatial ability, working memory) showed no significant influence. These findings demonstrate that task demands outweigh individual differences in shaping classification performance, highlighting the potential for task-adaptive navigation systems that dynamically adjust map complexity in response to real-time cognitive states.

Keywords

Cognitive Load, EEG, Machine Learning, Virtual Reality, Navigation, Human-Computer Interaction

1 Introduction

Spatial learning, which involves navigating through physical space, is a fundamental component of human cognition. It enables individuals to encode landmarks, form spatial representations, and remember directions. Effective spatial learning is essential in daily life, whether navigating urban centers, commuting through public transport systems, or exploring unfamiliar places while traveling. However, in today's digital society, people increasingly rely on GPS-based mobile maps to guide their navigation. Although such systems offer efficiency and convenience, a growing body of research suggests that they negatively affect spatial learning performance [1–7]. By providing turn-by-turn instructions, mobile maps reduce engagement with environmental cues and lead to weaker

acquisition of survey knowledge and landmark memory [4, 5]. Understanding the cognitive processes involved in map-assisted navigation is therefore critical for designing adaptive human–machine systems that support both efficient wayfinding and spatial learning.

Navigation can be conceptualized as a dual-task paradigm: locomotion through the environment is cognitively straightforward, while the secondary task of constructing an internal representation of the environment is demanding [8]. Landmarks are central to this process, serving as salient environmental anchors that structure mental representations and support route memory and decision making [9]. However, processing landmarks requires cognitive resources, and excessive numbers can impose excessive mental load. Research indicates that while a moderate number of landmarks may aid learning, higher numbers can overwhelm users, increasing intrinsic cognitive load and impairing performance [10, 11].

Empirical work confirms this trade-off [11, 12], finding that spatial learning performance (landmark recognition and route memory) improved when the number of on-map landmarks increased from three to five, but did not improve further with seven landmarks. This behavioral evidence for an optimal landmark 'sweet spot' was supported by EEG data showing increased frontal theta Event-Related Synchronization (ERS) and parieto-occipital P3 amplitude in the 7-landmark condition compared to the 5-landmark one, consistent with elevated cognitive load. These findings demonstrate that while landmarks are indispensable, excessive information can undermine spatial learning. Nevertheless, the analysis was limited to group-level comparisons of predefined biomarkers [13–17]. Thus, the relationship between moment-to-moment cognitive load states and spatial learning outcomes remains unquantified at the individual level, and the potential of a richer, multivariate neural signature to discriminate subtler states of cognitive overload remains unexplored.

These observations align with cognitive load theory, which defines load as the mental effort required for information processing [18]. In navigation, locomotion is relatively automatic, while wayfinding demands higher-level cognitive operations such as attention, working memory, and spatial reasoning. This dual demand increases cognitive load, particularly when users must attend to environmental cues while simultaneously consulting a mobile map. The design of navigation systems can either mitigate or exacerbate this load. Conventional GPS-based maps focus attention on

arXiv:2509.14056v1 [cs.HC] 17 Sep 2025

the device, increasing extraneous load and reducing opportunities for spatial learning [4, 19]. By contrast, including landmarks in digital maps can improve spatial knowledge acquisition, but excessive or poorly structured landmark information risks overloading the user [11, 20]. Thus, striking the right balance is essential for designing navigation adaptive systems that support both efficient wayfinding and long-term spatial learning.

Traditional methods for assessing cognitive load in navigation have relied on dual-task paradigms and self-report instruments such as the NASA-TLX [21]. However, dual-task methods can interfere with natural navigation, while self-reports offer only retrospective and subjective judgments that lack temporal resolution [20, 22]. These limitations highlight the need for direct, unobtrusive, and temporally precise measures of cognitive load in ecological navigation tasks. Building on these challenges, quantifying cognitive load during navigation in real-world settings is difficult due to uncontrollable factors such as traffic, weather, and individual behaviors (e.g., walking speed) [23, 24]. VR offers a powerful methodological solution by combining ecological validity with experimental control. Immersive 3D environments allow researchers to replicate naturalistic navigation while systematically manipulating variables such as the number of landmarks on a map—conditions that are difficult to achieve in the real world [25, 26]. Furthermore, VR is fully compatible with neurophysiological recording methods such as EEG, enabling the synchronous capture of brain activity and behavior during active navigation [15, 24]. Thus, a VR-based paradigm provides an effective platform for investigating the cognitive load of landmark processing.

Recent work has investigated diverse physiological measures for cognitive load detection, such as hearth rate variability (HRV), Eye-Tracking (ET), functional Near-Infra Red Spectroscopy (fNIRS), Electrodermal activity (EDA), and EEG [27–31]. Reliable real-time assessment of cognitive load is particularly important in Human-Computer Interaction (HCI), education, healthcare, and aviation [32–35]. Among these modalities, EEG is especially suitable because it captures brain activity at millisecond resolution, enabling the tracking rapid changes in mental effort without disrupting ongoing behavior. Well-established neural signatures include increases in frontal theta and decreases in parietal alpha power [13–15, 36]. Building on these foundations, EEG has been applied across domains such as education, driving, and aviation, demonstrating its value for real-time workload monitoring [11, 31, 33, 37–41]. In navigation, it further enables the tracking of cognitive demand fluctuations as users interact with mobile maps.

Advances in Machine Learning (ML) have further expand EEG's potential. By combining multivariate features such as spectral band power, temporal dynamics, and statistical descriptors, classifiers including Logistic Regression (LR), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Random Forests (RF), Spiking Neural Network (SNN), Extreme Gradient Boosting (XGBoost) and Artificial Neural Network (ANN) can distinguish between different cognitive states [31, 38–40, 42–46]. These methods have achieved promising accuracy in controlled paradigms such as the n-back and go/no-go tasks, as well as in semi-naturalistic yet controlled settings such as driving or reading, demonstrating the feasibility of EEG-based cognitive load classification. However, their application to ecologically valid navigation tasks remains limited.

A central pursuit in HCI research is the design of adaptive systems that sense user states and adjust interfaces in real time to improve usability, preventing overload, and enhance task performance [34, 47, 48]. Early approach followed a user-adaptive paradigm, tailoring interaction to relatively stable individual profiles such as demographics, expertise, user preferences and cognitive ability [32, 35, 36, 47–49]. More recent frameworks emphasize task- or state-adaptive approaches, which dynamically adjust interfaces in response to momentary task demands or user states [34, 50, 51]. Physiological adaptation systems exemplify this shift, modifying task difficulty based on affective or workload classification [49].

In navigation, it remains unclear whether EEG-based classification is driven more by stable user traits or by the dynamic complexity of the task environment. Resolving this distinction is essential for designing adaptive systems that balance efficiency with spatial learning in realistic contexts.

This work builds on a VR navigation paradigm developed for studying cognitive load [12], from which the dataset was collected. Previous studies of digital map navigation analyzed EEG spectral power and Event-Related Potential (ERP)s [11, 12], but emphasized group-level effects. To our knowledge, no prior research has applied ML to detect cognitive load from EEG in digital map-based navigation. Here, we address this gap by evaluating classifiers across EEG channel subsets, moving beyond group-level averages toward trial-by-trial decoding of load induced by varying landmark quantities.

Our analysis addresses two research questions:

- RQ1. Can EEG-based ML classifiers reliably distinguish levels of cognitive load during VR navigation?
- RQ2. Is variation in classifier performance better explained by individual traits (e.g., age, gender, working memory, spatial ability, perspective-taking) or by task complexity?

These comparisons directly inform the debate between user-adaptive and state-adaptive approaches in HCI. Methodologically, the study establishes a pipeline for EEG-based load classification in realistic navigation contexts. Conceptually, it provides empirical evidence on whether adaptive navigation systems should prioritize stable user profiles or dynamic detection of cognitive states.

2 Methods

2.1 Dataset and Experimental Paradigm

Participants: Forty-seven participants (29 female, 18 male; age range = 18–35 years, M = 25.6, SD = 4.09) were recruited for a study on spatial learning and cognitive load during virtual navigation [12]. All provided written informed consent in accordance with ethical guidelines from the University's Ethics Board, the Swiss Psychological Society, and the American Psychological Association. Each participant received compensation of 30 CHF.
Individuals with a history of neurological or psychiatric disorders were excluded. One participant (ID 51) was removed due to incomplete data, yielding a final sample of 46. Full details of the dataset and behavioral results are reported in the precursor study [12].

Participants navigated predefined routes in three distinct virtual European-style cities, designed in ArcGIS CityEngine 2018.0 and displayed in a stereoscopic Cave Automatic Virtual Environment (CAVE) system using Unity 2018.4 LTS (Fig. 1). Participants
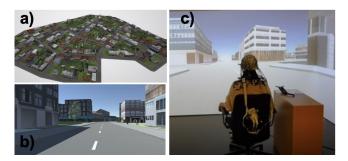
Figure 1: Virtual navigation setup. Experimental setup. (a) Bird's-eye view of a virtual city layout used for navigation. (b) First-person street-level perspective during navigation. (c) Participant seated in a three-wall CAVE environment while wearing an EEG cap. The CAVE projects high-resolution 3D images onto three surrounding walls, with the display perspective continuously updated based on head position and orientation, providing an immersive navigation experience during EEG recording. [11].

navigated from a first-person perspective while seated and instrumented for EEG.

In this work, we analyzes EEG data from a map-assisted navigation task. A within-participants design was used to examine how the number of landmarks (3, 5, or 7) displayed on a mobile map influenced cognitive load. These quantities were chosen visuospatial working memory research [52–55], which estimates a core capacity of about four items, with higher limits for meaningful real-world objects such as landmarks [54, 56, 57]. This design allowed us to probe low (3), medium/high (5), and high/overload (7) cognitive load conditions. The precursor study confirmed the validity of this manipulation: the 7-landmark condition elicited significantly greater cognitive load, reflected in increased frontal theta power and P3 amplitude relative to the 3- and 5-landmark conditions.

Task and Procedure. Participants were instructed to reach a destination as quickly as possible using a rotating map displayed on the central screen. The map provided turn-by-turn instructions and indicated the participant's current location and heading. It appeared for 5-second intervals at 17 predefined points along each route (before and after intersections, and on straight segments) (Fig. 2). During these intervals, the virtual environment was hidden and navigation was paused, simulating the real-world behavior of stopping to consult a mobile device. This design required participants to rely on memory for route continuation and landmark learning between map appearances.

Landmark Manipulation (Independent Variable). The independent variable was the number of landmarks displayed on the map. Visually salient buildings at intersections were selected and rendered in 3D as landmarks, based on the criteria of persistence, salience, and informativeness [58]. Landmark positions were chosen to ensure equal spatial distribution along the route. Each participant experienced all three conditions, counterbalanced across cities to control order effects:
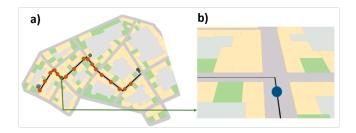

Figure 2: Map pop-up procedure. a) 17 pop-up locations (red dots) along the predefined route. b) Example pop-up map indicating the participant's current location (blue dot) and the upcoming route segment. The number of displayed landmark (3/5/7) varied according to assigned condition.

- 3-landmark condition
- 5-landmark condition
- 7-landmark condition

On each pop-up map, either 3, 5, or 7 landmarks were displayed, as shown in Fig. 3.

After navigating each city, participants' spatial knowledge was assessed through a battery of tests including landmark recognition, route direction, and a Judgement of Relative Direction (JRD) task. The analysis of these behavioral measures is reported in the precursor study [11] and is not the focus of the current analysis. Responses were collected using a 3D responding and pointing device (WorldViz Inc, USA).

Cognitive Measures. Individual differences in spatial navigation ability and visuospatial cognitive capacity were assessed using established instruments:

- The Santa Barbara Sense of Direction Scale (SBSOD) [59], a self-report questionnaire measuring spatial orientation abilities and navigational style.
- The Perspective Taking/Spatial Orientation Test (PTSOT) [60], which assesses the ability to imagine and orient from different spatial perspectives, with performance scored as angular error (in degrees).
- The Corsi Block-Tapping Task [61], a visual-spatial test of working memory capacity, scored as the maximum span length (partial score) a participant can correctly recall.

2.2 EEG Acquisition and Preprocessing

EEG was recorded using a 64-channel LiveAmp system (Brain Products GmbH) with active electrodes arranged in an extended 10–20 montage. The signal was referenced to FCz, grounded at Fpz, and sampled at 500 Hz. Impedances were kept below 10 kΩ. Data were streamed wirelessly via a UBT21 Bluetooth adapter and synchronized with navigation events (e.g., map onsets/offsets) using inter-process communication under Windows.

EEG data were preprocessed and analyzed using MNE-Python. Signals were bandpass filtered between 0.5–45 Hz and notch filtered at 50 Hz and harmonics to remove the line noise. Bad channels were identified and excluded, and data were re-referenced to the average reference. Excluded channels were reconstructed via interpolation. Epochs were extracted from −1 s to +5 s around each "showMap"
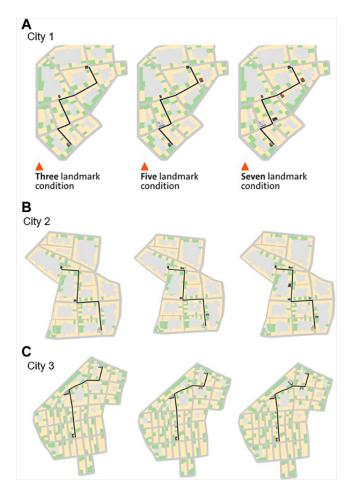
Figure 3: Landmark-density conditions. Example mobile maps for the three conditions (3, 5, 7 landmarks). Conditions were counterbalanced across cities.

event, with a baseline correction applied over −1 s to 0 s interval. Indipendent Component Analysis (ICA) was performed using the FastICA algorithm [62], retaining components explaining 99% of variance. Artifactual components (ocular, muscular, cardiac, line noise, channel noise) were automatically identified with ICLabel and subsequently removed.

3 Feature Extraction

From the signal a total of 143 features were extracted. The decision to extract a broad set of features from temporal, spectral, and temporal–spectral domains was motivated by prior work [38] showing that different signal characteristics capture complementary aspects of cognitive processing. Temporal features reflect variability and dynamics of ongoing neural activity, spectral features highlight oscillatory processes such as theta and alpha rhythms linked to attention and working memory, and temporal–spectral features capture transient changes that often occur during demanding cognitive states. As Wittke et al. [38] argue, using a rich feature set

maximizes the likelihood of identifying discriminative neural markers in ecologically valid tasks, where noise and inter-individual variability may otherwise obscure relevant patterns. To mitigate risks of overfitting in such high-dimensional spaces, we adopted ML models with built-in regularization alongside cross-validation approach. We used ML models that have been repeatedly shown to handle complex physiological data effectively [46] and provide interpretable insights into feature contributions. In addition, we compared multiple model families, including linear classifiers and neural architectures to ensure that our results were not biased by the inductive biases of any single approach. The use of nested cross-validation further ensured that hyperparameter tuning was separated from performance estimation, which is a critical step for generalization in HCI contexts where end-user deployment requires robustness across unseen individuals and environments.

Temporal features (13 features): Channel-averaged statistics derived from the raw EEG, including amplitude measures (mean, Standard Deviation (SD), minimum, maximum, peak), signal complexity metrics (outlier ratio, root mean square, zero-crossing rate), distribution descriptors (skewness, kurtosis), and Hjorth parameters (activity, mobility, complexity).

Spectral features (22 features): Band power was computed using Welch's method (2-second Hanning windows, 50% overlap) to capture five canonical frequency ranges: δ (0.5–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–30 Hz), and γ (30–45 Hz). Derived features included global inter-hemispheric asymmetry indices (δ_asy, θ_asy, α_asy, β_asy, γ_asy), regional asymmetries (e.g., frontal α_asy, frontotemporal indices), and within-channel ratios such as α/θ, θ/α, α/β, θ/β, $(\theta + \alpha)/\beta$, and $(\theta + \alpha)/(\alpha + \beta)$. Asymmetry was defined as:

$$\text{Asymmetry} = \ln(P_{\text{right}} + \epsilon) - \ln(P_{\text{left}} + \epsilon),$$

with $\epsilon = 10^{-12}$ ensuring numerical stability.

Temporal–spectral features (108 features): To capture the spatial distribution of the spectral power, the band power and ratios per channel were aggregated within each epoch. For each of nine spectral metrics (δ, θ, α, β, γ, α/β, θ/β, $(\theta + \alpha)/\beta$, $(\theta + \alpha)/(\alpha + \beta)$), we computed 12 statistical descriptors across channels: mean, SD, minimum, maximum, median, Interquartile Range (IQR), coefficient of variation, skewness, kurtosis, Shannon entropy, absolute peak value and outlier ratio (fraction of channels exceeding 5 SD).

The features were extracted using a sliding window approach (2 s duration with 50% overlap) within each epoch, following methodologies established in recent literature [31, 38] while extending traditional feature sets with enhanced frequency domain and spatial aggregation metrics.

3.1 Machine Learning Pipeline

Classifiers and Hyperparameter Tuning. We evaluated five supervised classifiers: LR, SVM, RF, XGBoost, and Multilayer Perceptron (MLP). Before training, all features were standardized using `StandardScaler`.

For each model, the hyperparameters were optimized via a grid search within a nested cross-validation framework. The following parameter spaces were explored:

- **LR**: Regularization strength (C), penalty type (l1, l2), solver.
- **SVM**: Kernel (linear, rbf, poly), C, gamma, degree.

- **RF**: Number of estimators (n_estimators), maximum tree depth (max_depth).
- **XGBoost**: max_depth, learning_rate, subsample, colsample_bytree, reg_lambda.
- **MLP**: Hidden layer sizes, activation function, L2 regularization (alpha).

Training and Evaluation Procedure. A participant-wise nested stratified cross-validation was implemented to ensure robust generalizability and to prevent information leakage. For each participant, the available epochs were partitioned using an outer loop of 2 to 5 times (depending on the balance of the class). One fold served as the test set, while the remaining folds were reserved for training and hyperparameter optimization.

Within each outer training set, an inner cross-validation loop (also 2 to 5 folds) was conducted to select the optimal hyperparameters from the model-specific grids. The best configuration was then retrained on the entire outer training set and evaluated in the corresponding outer test set. This procedure was repeated across all folds, ensuring unbiased estimation of model performance. We considered four classification tasks:

(1) Multiclass: 3 vs. 5 vs. 7 landmarks,
(2) Binary: 3 vs. 5 landmarks,
(3) Binary: 3 vs. 7 landmarks,
(4) Binary: 5 vs. 7 landmarks.

In summary, this nested cross-validation framework provided an unbiased evaluation of each classifier's ability to distinguish cognitive load levels under varying landmark conditions, while maintaining separation of training and test data within each participant. Unlike prior EEG studies in navigation, which relied on aggregated statistical tests, our approach uses supervised ML to decode cognitive load on a per-trial basis.

Performance was assessed using multiple metrics: overall mean accuracy, macro-averaged F1-score, and per-class precision, recall, and F1-scores. Results were aggregated across outer folds to report mean and SD for each metric. In addition, averaged maximum accuracy across outer folds (cv_max) was recorded for each participant-model-task combination to capture potential upper-bound performance.

The macro-averaged F1-score was employed because it provides a more robust evaluation than accuracy in the presence of potential class imbalance across cross-validation folds. Unlike accuracy, which weights all predictions equally, the Macro F1 calculates the F1-score (the harmonic mean of precision and recall) independently for each class and then takes the arithmetic mean. This ensures that each class contributes equally to the final metric, preventing majority classes from dominating the performance assessment and providing a more comprehensive measure of a model's ability to distinguish between all cognitive states.

Finally, to investigate the contribution of electrode coverage to classification performance, we repeated the full ML pipeline using three different sets of EEG channels:

- All channels: full electrode montage after preprocessing.
- Frontal subset: electrodes over the frontal lobe, capturing activity associated with executive and working memory processes.
- Frontal–parietal subset: electrodes spanning both frontal and parietal cortices, targeting regions implicated in attentional control and working memory load.

To statistically assess differences between channel subsets, we compared paired participant-level scores using both paired t-tests and Wilcoxon signed-rank tests. Analyses were conducted separately for each task contrast (3-class classification, 3 vs. 5, 3 vs. 7, 5 vs. 7). Multiple testing was accounted for by reporting both p-values and effect sizes.

3.2 Analysis of Individual Differences in Classification Performance

To examine whether demographic and cognitive factors influenced which binary classification task yielded the best performance for each participant, we conducted a series of statistical analyses. Normality of continuous variables was assessed using the Shapiro–Wilk test.

In summary, we performed the following analyses:

(1) χ^2 test of independence between gender and optimal task classification
(2) One-way ANOVAs for normally distributed variables across task groups (with gender as a covariate)
(3) Kruskal–Wallis tests for non-normally distributed variables
(4) Correlation analyses between classification accuracy and continuous variables (Pearson, gender-weighted, and partial correlations controlling for gender)
(5) Multinomial logistic regression predicting optimal task classification

Group Comparison Tests. We conducted parametric (ANOVA) or non-parametric (Kruskal-Wallis) tests based on data normality to examine differences in cognitive measures across participants grouped by their optimal classification task. The ANOVA model included task as the independent variable and cognitive scores as dependent variables, with gender included as a covariate.

Data Preparation and Weighting. To address gender imbalance in the sample (28 female, 18 male participants), we calculated gender weights for each participant using:

$$w_g = \frac{N_{\text{total}}}{2 \times N_g} \tag{1}$$

where N_{total} represents the total sample size and N_g represents the number of participants of gender g. These weights were applied in subsequent weighted analyses to ensure balanced representation.

Correlation Analyses. We computed both regular Pearson correlations and gender-weighted correlations between classification accuracy and continuous variables of interest (age, SBSOD spatial ability score, PTSOT perspective-taking error, and Corsi working memory score). Weighted correlations were calculated using the DescrStatsW class from the statsmodels package which incorporates case weights in correlation computation.

For each continuous variable, we also calculated partial correlations with classification accuracy while controlling for gender

effects. The partial correlation coefficient $r_{12.3}$ was computed using:

$$r_{12.3} = \frac{r_{12} - r_{13} \times r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad (2)$$

where r_{12} represents the correlation between accuracy and the variable of interest, r_{13} represents the correlation between accuracy and gender, and r_{23} represents the correlation between the variable of interest and gender.

Multinomial Logistic Regression. We implemented a multinomial logistic regression model to predict participants' optimal classification task:

$$\text{Task} \sim \text{Age} + \text{Gender} + \text{SBSOD} + \text{PTSOT error} + \text{Corsi score} \qquad (3)$$
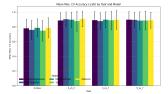
Gender was encoded numerically (Male=1, Female=0), and task categories were encoded using label encoding. The model was fit using maximum likelihood estimation. Model performance was evaluated using classification accuracy, confusion matrices, and comparison against null accuracy.
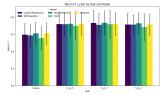
All analyses were conducted using Python 3.9 with statsmodels, scikit-learn, and SciPy libraries. Statistical significance was evaluated at $\alpha = 0.05$, and all tests were two-tailed.

4 Results

4.1 Model Performance Across Classification Tasks

The nested cross-validation evaluation revealed distinct performance patterns across the four classification tasks and five machine learning models. Table 1 summarizes the mean maximum cross validation accuracy (`cv_max`) for each model-task combination, representing the upper bound of achievable performance across cross-validation folds.

(a) Mean maximum CV accuracy across tasks and models.

(b) Macro F1 scores across tasks and models.

Figure 4: Comparison of classification performance across tasks and models. (a) Mean maximum cross validation accuracy across participants, (b) macro F1 scores. Error bars represent standard deviations.

The nested cross-validation results showed a consistent performance hierarchy across models (Fig. 4). Binary classification tasks achieved higher accuracy than the multiclass condition, with the 3 vs. 7 landmarks task yielding the highest performance. Maximum cross-validation accuracy reached 90.8% for the best model–task combination (Fig. 4a). Macro F1 scores exhibited the same pattern (Fig. 4b).

Table 1: Maximum Attainable Performance Across Classification Tasks (Mean ± SD, %)

Task and Model	Accuracy (mean `cv_max`, %)	Macro F1
3-class classification		
XGBoost	78.7% ± 13.7%	61.2% ± 12.9%
RF	78.6% ± 12.3%	61.2% ± 12.8%
LR	77.9% ± 12.6%	59.5% ± 13.3%
MLP	75.8% ± 13.1%	58.7% ± 13.1%
SVM	74.7% ± 14.2%	55.7% ± 13.4%
3 vs. 5 classification		
XGBoost	90.8% ± 10.4%	72.5% ± 13.4%
MLP	90.4% ± 10.0%	71.9% ± 13.7%
RF	89.6% ± 10.5%	72.5% ± 13.9%
LR	88.7% ± 10.8%	72.1% ± 14.3%
SVM	87.8% ± 13.6%	70.0% ± 16.6%
3 vs. 7 classification		
RF	89.6% ± 10.9%	73.6% ± 14.1%
XGBoost	89.5% ± 12.9%	72.0% ± 13.5%
LR	89.0% ± 12.1%	73.3% ± 14.7%
SVM	88.8% ± 12.5%	71.8% ± 16.3%
MLP	87.8% ± 11.4%	71.1% ± 13.7%
5 vs. 7 classification		
LR	89.8% ± 11.1%	71.5% ± 13.6%
MLP	89.5% ± 10.4%	71.5% ± 11.1%
XGBoost	89.1% ± 10.9%	71.4% ± 12.4%
SVM	88.7% ± 12.8%	68.6% ± 15.8%
RF	88.5% ± 12.2%	72.9% ± 12.2%

Key Findings. The `cv_max` metric, representing the maximum accuracy across outer folds, indicated that individual participants could achieve higher performance (up to 89-91% accuracy) than the average cross-validated results, suggesting significant potential for optimal model configuration though participant-specific factors may still influence optimal model performance. Moreover, the SD across folds were typically between 0.10 to 0.15, indicating considerable variability in performance across different data partitions.

Classification was strongest for binary tasks, peaking at 90.8% for 3 vs. 5 landmarks, but performance remained robust in the more complex 3-class problem (78.7%). Thus, while finer gradations of load are more challenging to distinguish, they remain feasible. Across models, tree-based approaches (RF and XGBoost) performed best, followed by LR, MLP, and SVM.

To further examine whether such performance patterns related to individual traits, we assessed distributions of participant measures. Shapiro–Wilk tests indicated that all variables except PTSOT error were normally distributed across task groups ($p > 0.05$), whereas PTSOT error significantly deviated from normality ($p < 0.05$). Analysis of the 46 participants further revealed a comparable distribution of optimal binary classification types: 18 participants (39.1%) performed best on the 3 vs. 5 landmarks task, 14 (30.4%) on the 3 vs. 7 task, and 14 (30.4%) on the 5 vs. 7 task.

Analysis of Demographic and Cognitive Variables. One-way ANOVAs controlling for gender showed no significant differences in age ($p = 0.769$), spatial ability ($p = 0.485$), or working memory capacity ($p = 0.465$) across participants grouped by their optimal classification task. Non-parametric Kruskal–Wallis test revealed no significant differences in perspective-taking error ($p = 0.954$).

A marginal association was observed between gender and optimal task, $\chi^2(2, N = 46) = 5.948$, $p = 0.051$. Female participants more often showed optimal performance on the 3 vs. 5 task (50.0%), whereas male participants more often favored the 3 vs. 7 task (50.0%).

Descriptive statistics are reported in Table 2 and inferential results are reported in Table 3.

Table 2: Descriptive statistics (mean ± SD) for demographic and cognitive variables by optimal classification task

Variable	3 vs. 5 (n=18)	3 vs. 7 (n=14)	5 vs. 7 (n=14)	p
Age	24.94 ± 4.72	25.36 ± 3.27	26.00 ± 3.90	0.769
SBSOD	4.91 ± 0.96	4.46 ± 1.31	4.76 ± 0.85	0.485
PTSOT	26.33 ± 21.20	27.18 ± 23.84	21.29 ± 11.22	0.954
Corsi	4.33 ± 0.61	4.12 ± 0.50	4.13 ± 0.55	0.465

Table 3: Inferential test results summary for demographic and cognitive variables across task groups

Analysis	Result
Gender–Task Association	$\chi^2(2, N = 46) = 5.948, p = 0.051$
Age by Task	$F(2, 43) = 0.264, p = 0.769$
SBSOD by Task	$F(2, 43) = 0.736, p = 0.485$
PTSOT Error by Task	$\chi^2(2, N = 46) = 0.095, p = 0.954$
Corsi Score by Task	$F(2, 43) = 0.780, p = 0.465$

Correlational and Predictive Analyses. Classification accuracy showed no significant correlations with age ($r = 0.124$, $p = 0.411$), spatial ability ($r = -0.048$, $p = 0.752$), perspective-taking error ($r = -0.171$, $p = 0.255$), or working memory ($r = 0.003$, $p = 0.984$). These null results remained consistent when applying gender weights, conducting gender-stratified analyses, and controlling for gender via partial correlations (all $|r| < .20$).

A multinomial logistic regression model predicting optimal task classification achieved limited accuracy (47.8%), only marginally exceeding the null accuracy rate (39.1%). No demographic or cognitive variables emerged as significant predictors in the model.

Our analyses revealed a consistent pattern: demographic and cognitive variables showed no significant predictive power for optimal classification performance. A marginal trend with gender was observed ($p = 0.051$), but overall, classifier performance was shaped primarily by task demands rather than stable participant traits.

Fig. 5 illustrates these relationship. Panels a)–e) show gender-stratified boxplots for accuracy, spatial ability, working memory, age, and perspective-taking error across task groups. Although inferential tests did not reveal significant group differences, several weak, non-significant trends are visible: a) men showed slightly higher median accuracies in the 3 vs. 7 and 5 vs. 7 tasks; b) spatial ability (SBSOD) indicated a small male advantage in the 3 vs. 7 group; c) working memory fluctuated inconsistently across tasks and genders; d) the 5 vs. 7 group skewed marginally older; and e) males tended to have lower PTSOT error in the 3 vs. 7 and 5 vs. 7 tasks, though with high variability. Panel f) shows the confusion matrix of the multinomial logistic regression, which reached only modest accuracy (≈48%), slightly above the null baseline, with most correct classifications in the 3 vs. 5 group and frequent confusions between 5 vs. 7 and 3 vs. 5.

4.2 The Impact of Electrode Coverage on Classification Performance

We compared models trained on all 64 channels, a frontal-only subset, and a frontal–parietal subset. Across metrics and classifiers, the full montage consistently yielded the highest performance (Fig. 6). Restricting input to the frontal subset produced only minor accuracy losses ($\Delta \approx$ 1–2%), while the frontal–parietal subset showed larger decrements ($\Delta \approx$ 2–3%), particularly in the three-class condition where most differences reached statistical significance (p < 0.01; Fig. 7). For binary contrasts (3 vs. 7, 5 vs. 7), the frontal subset achieved comparable results to the all-channel baseline (p > 0.7), whereas the frontal–parietal configuration again underperformed. Interestingly, extending coverage to parietal sites did not provide systematic benefits and in some cases even reduced performance relative to frontal-only input.

These effects were stable across classifiers, with RF and XGBoost generally achieving the highest absolute scores, but following the same ranking of channel sets.

5 Discussion

Addressing RQ1, our results demonstrate that EEG-based classifiers can reliably distinguish levels of cognitive load, with accuracies reaching 78.7% for three-class and up to 90.8% for binary contrasts, underscoring the feasibility of EEG-based workload detection in ecologically valid but controlled VR tasks. These values are competitive with, and in some cases exceed, performance reported in other domains. For example, simulated driving studies report similar accuracy ranges: Wang et al. [63] achieved ~90% for two-class classification using deep neural networks, and Di Flumeri et al. [33] reported ~0.75 AUC for three-class classification when combining EEG and ET. Laboratory paradigms such as the n-back task typically yield accuracies in the 70–80% range for three-class classification task (e.g., 77.2% in [64]; 74.9% in [65]). By contrast, more unconstrained scenarios show marked reductions: Wittke et al. [38] reported only ~60% F1 for digital textbook reading using a similar feature-extraction and ML pipeline. Together, these comparisons indicate that structured yet dynamic environments, such as VR navigation or driving simulators, offer the balance of ecological validity and experimental control needed to elicit robust neural signatures of cognitive load.

Our findings addressing RQ2 provide clear evidence that EEG-based cognitive load classification during navigation is primarily task-driven rather than user-driven. Classifiers decoded load across
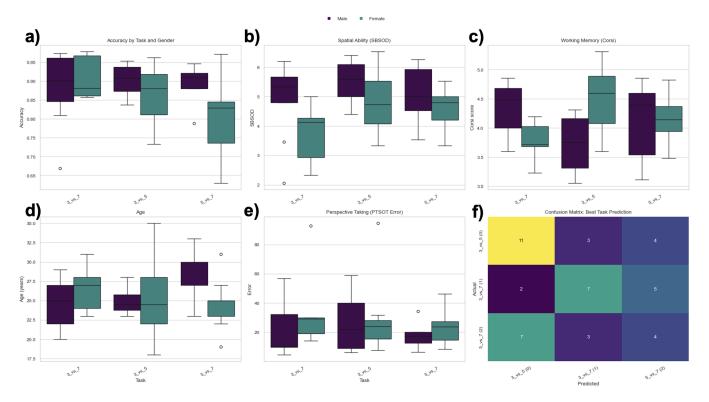
Figure 5: Visualization of demographic and cognitive variables across task groups and genders, and confusion matrix of multinomial logistic regression. Boxplots show distributions of (a) classification accuracy, (b) spatial ability SBSOD, (c) working memory (Corsi score), (d) age, and (e) perspective-taking error PTSOT across gender and optimal classification task. These plots corroborate the statistical findings: group differences are small and largely overlapping, and model predictions only marginally exceeded the null baseline (f) Confusion matrix illustrates predictive performance of the multinomial logistic regression model.
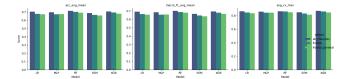

Figure 6: Model performance across electrode subsets. Bars show mean accuracy, macro-F1, and cv_max for each classifier using all channels, frontal, and frontal–parietal presets. This figure highlights the relative performance hierarchy of channel sets across metrics.

Figure 7: Statistical comparison of presets by task. Each panel summarizes pairwise tests across tasks and metrics: *all vs. frontal* (left), *all vs. frontal–parietal* (middle), and *frontal–parietal vs. frontal* (right). Cells encode outcomes (ns, trend, sig); the all-channels montage significantly outperforms frontal–parietal in most 3-class/3vs5 metrics, while frontal often remains comparable to all in the binary tasks.

landmark conditions with accuracies well above chance (Table 1), yet performance was systematically determined by task complexity rather than individual characteristics. Neither demographic (age, gender) nor cognitive factors (working memory, spatial ability, perspective-taking) predicted classification outcomes (Table 3). This suggests that the dominant source of variance in neural signatures of cognitive load arises from the task environment itself rather than from stable user profiles.

Cognitive load during map-assisted VR navigation increases with the number of displayed landmarks: learning improves from 3 to 5

landmarks without additional load, but 7 landmarks elicit higher frontal theta and parietal P3 responses, reflecting cognitive overload without learning benefits [12]. Our findings extend this literature by showing that these neural responses are not only present but also discriminable at the level of individual trials. This provides direct

support for the "inefficient overload" state described in earlier work, where adding landmarks increases neural effort without producing behavioral gains. Critically, classification accuracy exceeded 83% even for the subtle 5 vs. 7 landmark contrast, a condition that is behaviorally indistinguishable. This demonstrates that overload is reflected in distributed, multivariate neural patterns spanning spectral, temporal, and time–frequency domains, rather than in isolated biomarkers. Whereas traditional univariate analyses captured only the most prominent components (e.g., frontal theta, parietal P3), our feature-rich approach revealed finer-grained, spatially distributed signatures that more comprehensively characterize overload states.

Unlike prior workload studies in education or driving, our work directly addresses digital navigation interfaces, an important HCI application where adaptive interaction design can immediately benefit from reliable cognitive load detection. From this perspective, these results move the field from descriptive group-level analyses toward predictive, trial-level modeling of cognitive state. By showing that EEG can track load dynamically in real time, our work lays a foundation for closed-loop neuro-adaptive systems. Such systems could, for example, regulate the number or salience of landmarks displayed on a digital map to prevent overload while maintaining wayfinding efficiency. Importantly, classifier performance was largely comparable across different algorithms (LR, MLP, RF, SVM, XGBoost) once task was controlled and with a rich set of features, indicating that task demands outweighed model choice in shaping performance. This finding suggests that designing adaptive systems should prioritize optimizing task conditions rather than seeking marginal gains from algorithmic tuning.

Electrode subset analyses further support the feasibility of lightweight, applied EEG systems. Frontal electrodes alone were sufficient to capture the neural signatures of cognitive load, particularly in binary contrasts, while adding parietal sites unexpectedly reduced performance. This may reflect additional noise or inter-individual variability introduced by parietal signals not directly aligned with load-sensitive activity. Although the full 64-channel montage produced the highest overall accuracies, the frontal-only configuration yielded statistically comparable performance for critical contrasts (e.g., high vs. overload). This is encouraging for applied neuroergonomics, as it demonstrates that compact, wearable EEG systems could achieve practical utility in real-world navigation contexts.

Several limitations should be acknowledged. First, the participant pool consisted of young, highly educated adults, limiting generalizability to older or more diverse populations [66, 67]. Future work should examine whether similar cognitive capacity limits emerge in older adults, who may experience reduced working memory and slower attentional shifts, or in populations with different educational and cultural backgrounds. Second, the VR environment was constrained to medium-length, grid-like routes with track-up map orientation; results may differ in more organic city layouts, under north-up map orientations, or when participants navigate longer and more complex journeys. Third, we examined load during initial encoding in a novel environment, whereas load dynamics likely evolve with repeated exposure, consolidation, and retrieval.

Future research should address these limitations by testing more heterogeneous participant groups and more varied navigation environments. Another promising direction is multimodal integration. Prior studies show that combining EEG with complementary measures such as ET, EDA, or HRV substantially improves classification [28, 31, 40, 68, 69]. Recording gaze behavior alongside EEG could, for example, link neural signatures of load with fixation patterns, saccades, and pupil responses, producing more comprehensive and reliable estimates. Experimental manipulations could also amplify cognitive demand by varying route complexity, landmark salience, or time pressure, or by targeting retrieval phases such as spatial memory recall and intersection decision-making. Ultimately, advancing toward closed-loop paradigms will be essential: systems that detect overload in real time and dynamically adapt map design (e.g., simplifying routes, modulating landmark density) will allow direct evaluation of whether adaptive interventions improve both usability and spatial learning.

6 Conclusion

In conclusion, this study demonstrates that EEG-based cognitive load classification during navigation is primarily task-driven rather than user-driven. To our knowledge, it provides the first evidence that EEG signals collected during digital map navigation can be reliably classified into load states, with accuracies up to 90.8% for optimal contrasts. Crucially, robust decoding was achievable even with reduced frontal montages, underscoring the feasibility of lightweight, portable systems for real-world use. These findings indicate that adaptive navigation aids should prioritize dynamic adjustment to task demands rather than static user profiling. Such real-time adaptation offers a pathway to preventing overload and enhancing spatial learning.

Acknowledgments

References

[1] Lama Dahmani and Veronique D. Bohbot. Habitual use of gps negatively impacts spatial memory during self-guided navigation. *Scientific Reports*, 10(1):1–14, 2020.

[2] Roger McKinlay. Technology: Use or lose our navigation skills. *Nature*, 531(7596):573, 2016.

[3] Avi Parush, Shahar Ahuvia, and Ido Erev. Degradation in spatial knowledge acquisition when using automatic navigation systems. In *International Conference on Spatial Information Theory*, pages 238–254, 2007.

[4] Toru Ishikawa. Satellite navigation and geospatial awareness: Long-term effects of using navigation tools on wayfinding and spatial orientation. *The Professional Geographer*, 71(2):197–209, 2019.

[5] Ian T. Ruginski, Sarah H. Creem-Regehr, Jeanine K. Stefanucci, and Elizabeth Cashdan. Gps use negatively affects environmental learning through spatial transformation abilities. *Journal of Environmental Psychology*, 64:12–20, 2019.

[6] Masaki Sugimoto, Takashi Kusumi, Noriyuki Nagata, and Tetsuya Ishikawa. Online mobile map effect: How smartphone map use impairs spatial memory. *Spatial Cognition and Computation*, 2021.

[7] Stefan Münzer, Hubert D. Zimmer, and Jörg Baus. Navigation assistance: A trade-off between wayfinding support and configural learning support. *Journal of Experimental Psychology: Applied*, 18(1):18–37, 2012.

[8] Daniel R. Montello. Navigation. In *The Cambridge Handbook of Visuospatial Thinking*, pages 257–294. Cambridge University Press, 2005.

[9] Kai-Florian Richter and Stephan Winter. *Landmarks: GIScience for Intelligent Services*. Springer International Publishing, 2014.

[10] John Sweller, Jeroen J. G. van Merriënboer, and Fred G. W. C. Paas. Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3):251–296, 1998.

[11] Bingjie Cheng. *Human Spatial Navigation in the Digital Era: Effects of Landmark Depiction on Mobile Maps on Navigators' Spatial Learning and Brain Activity During Assisted Navigation*. PhD thesis, University of Zurich, 2023.

[12] Bingjie Cheng et al. The effect of landmark visualization in mobile maps on brain activity during navigation: A virtual reality study. *Frontiers in Virtual Reality*, 3:981625, 2022.

[13] Wolfgang Klimesch. Eeg alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews*, 29(2-3):169–195, 1999.

[14] Thi Thu Nhi Do, Chin-Teng Lin, and Klaus Gramann. Human brain dynamics in active spatial navigation. *Scientific Reports*, 11(1):1–12, 2021.

[15] Alexandre Delaux, Jean-Baptiste de Saint Aubert, Stephen Ramanoël, Mathieu Bécu, Lars Gehrke, Marius Klug, Ricardo Chavarriaga, José-Alain Sahel, Klaus Gramann, and Angelo Arleo. Mobile brain/body imaging of landmark-based navigation with high-density eeg. *European Journal of Neuroscience*, 2021.

[16] Yu-Kai Wang, Tzyy-Ping Jung, and Chin-Teng Lin. Theta and alpha oscillations in attentional interaction during distracted driving. *Frontiers in Behavioral Neuroscience*, 12:3, 2018.

[17] Shuai Fu and Raja Parasuraman. Event-related potentials (erps) in neuroergonomics. In Raja Parasuraman and Matthew Rizzo, editors, *Neuroergonomics: The Brain at Work, Human Technology Interaction Series*, pages 32–50. Oxford University Press, New York, NY, 2006.

[18] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

[19] Aaron L. Gardony, Tad T. Brunyé, Caroline R. Mahoney, and Holly A. Taylor. How navigational aids impair spatial memory: Evidence for divided attention. *Spatial Cognition & Computation*, 13(4):319–350, 2013.

[20] Amy L. Griffin, Tumasch Reichenbacher, Hua Liao, Wangshu Wang, and Yinghui Cao. Cognitive issues of mobile map design and use. *Journal of Location Based Services*, 18(4):350–380, 2024.

[21] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, volume 52, pages 139–183. Elsevier, 1988.

[22] Bingjie Cheng, Enru Lin, Klaus Gramann, and Anna Wunderlich. Eye blink-related brain potentials during landmark-based navigation in virtual reality (short paper). In *15th International Conference on Spatial Information Theory (COSIT 2022)*, volume 240, pages 1–8. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

[23] Rudolph P. Darken and Barry Peterson. Spatial orientation, wayfinding, and representation. In *Handbook of Virtual Environments*, page 493–518. CRC Press, 2002.

[24] Klaus Gramann. Embodiment of spatial reference frames and individual differences in reference frame proclivity. *Spatial Cognition and Computation*, 13(1):1–25, 2013.

[25] Maria V. Sanchez-Vives and Mel Slater. From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, 6(4):332–339, 2005.

[26] Marius Klug, Stephanie Jeung, Anna Wunderlich, Lars Gehrke, Jochem Protzak, Zak Djebbara, Andreas Argubi-Wollesen, Bettina Wollesen, and Klaus Gramann. The bemobil pipeline for automated analyses of multimodal mobile brain and body imaging data. *BioRxiv*, 2022.

[27] Peizheng Wang, Robert Houghton, and Arnab Majumdar. Detecting and predicting pilot mental workload using heart rate variability: A systematic review. *Sensors*, 24(12):3723, 2024.

[28] T. Qin, W. Fias, N. Van de Weghe, and H. Huang. Recognition of map activities using eye tracking and eeg data. *International Journal of Geographical Information Science*, 2024.

[29] Şeniz Harputlu Aksu, Erman Çakıt, and Metin Dağdeviren. Mental workload assessment using machine learning techniques based on eeg and eye tracking data. *Applied Sciences*, 14(6):2282, 2024.

[30] Jiahui An, Pulkit Goyal, Andreas R. Luft, and Josef G. Schönhammer. Functional near-infrared spectroscopy short channel regression improves cortical activation estimates of working memory load. *Neurophotonics*, 12(3):035009, 2025.

[31] C. Anders, S. Moontaha, S. Real, et al. Unobtrusive measurement of cognitive load and physiological signals in uncontrolled environments. *Scientific Data*, 11:1000, 2024.

[32] Frédéric Dehais, Alex Lafont, Raphaëlle Roy, and Stephen Fairclough. A neuroergonomics approach to mental workload, engagement and human performance. *Frontiers in Neuroscience*, 14:268, 2020.

[33] Gianluca Di Flumeri, Gianluca Borghini, Pietro Aricò, Nicolina Sciaraffa, Paola Lanzi, Simone Pozzi, Valeria Vignali, Claudio Lantieri, Arianna Bichicchi, Andrea Simone, and Fabio Babiloni. Eeg-based mental workload neurometric to evaluate the impact of different traffic and road conditions in real driving settings. *Frontiers in Human Neuroscience*, 12:509, 2018.

[34] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. A survey on measuring cognitive workload in human-computer interaction. *ACM Computing Surveys*, 55(13s):39, 2023.

[35] Luca Longo. Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLOS ONE*, 13(8):e0199661, 2018.

[36] Alan Gevins and Michael E. Smith. Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science*, 4(1-2):113–131, 2003.

[37] Jahid Hassan, Md Shamim Reza, Syed Udoy Ahmed, Nazmul Haque Anik, and Md Obaydullah Khan. Eeg workload estimation and classification: A systematic review. *Journal of Neural Engineering*, 2024.

[38] Lennart Wittke, Peter El Hachem, Yuan Chen, Joseph Ollier, Bingjie Cheng, Shkurta Gashi, and Xiaoyu Zhang. Classifying cognitive load in digital textbook reading using electroencephalogram signals. In *2025 13th International Conference on Information and Education Technology (ICIET)*, pages 316–321, 2025.

[39] Jiahui An, Chonghao Cai, Olympia Gallou, Sara Irina Fabrikant, Giacomo Indiveri, and Elisa Donati. Neuromorphic Deployment of Spiking Neural Networks for Cognitive Load Classification in Air Traffic Control. In *2025 International Conference on Neuromorphic Systems (ICONS)*. IEEE Computer Society, 2025.

[40] Jiahui An, Sara Irina Fabrikant, Giacomo Indiveri, and Elisa Donati. Spiking neural networks for mental workload classification with a multimodal approach. In *2025 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8. IEEE Computer Society, 2025.

[41] Arun Balakrishna and Tom Gross. What humans should be thinking while driving: Method for integration of driver cognitive load information with map data. In *HCI International 2024 – Late Breaking Papers*, page 3–12. Springer-Verlag, 2024.

[42] Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara van Gog. Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4):425–438, 2010.

[43] Seung-Hyeon Oh, Yu-Ri Lee, and Hyoung-Nam Kim. A novel eeg feature extraction method using hjorth parameter. *International Journal of Electronics and Electrical Engineering*, 2(2):106–110, June 2014.

[44] Hojjat Adeli, Ziqin Zhou, and Nahid Dadmehr. Analysis of eeg records in an epileptic patient using wavelet transform. *Journal of Neuroscience Methods*, 123(1):69–87, February 2003.

[45] Joseph K. Nuamah and Younho Seong. Support vector machine (svm) classification of cognitive tasks based on electroencephalography (eeg) engagement index. *Brain-Computer Interfaces*, 5(1):1–12, 2018.

[46] Yi Ding, Yaqin Cao, Vincent G. Duffy, Yi Wang, and Xuefeng Zhang. Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning. *Ergonomics*, 63(7):896–908, 2020.

[47] Gerhard Fischer. User modeling in human–computer interaction. *User Modeling and User-Adapted Interaction*, 11(1–2):65–86, 2001.

[48] Anthony Jameson. Adaptive interfaces and agents. In Andrew Sears and Julie A. Jacko, editors, *The Human-Computer Interaction Handbook*, pages 305–330. Lawrence Erlbaum Associates, 2007.

[49] Ali Darzi, Asad Ali, Muhammad Awais, et al. Automated affect classification and task difficulty adaptation: A review. *International Journal of Human-Computer Studies*, 154:102673, 2021.

[50] Marcel Dubiel, Daniel Buschek, and Michael Sedlmair. A contextual framework for adaptive user interfaces. *arXiv preprint arXiv:2203.16882*, 2022.

[51] Sergio Yañez, Alexander Lex, Bahador Saket, et al. The state of the art in user-adaptive visualizations. *Computer Graphics Forum*, 44(1):123–147, 2025.

[52] Steven J. Luck and Edward K. Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281, 1997.

[53] Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114, 2001.

[54] Alan Baddeley. Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, 4(10):829–839, 2003.

[55] George A. Alvarez and Patrick Cavanagh. The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, 15(2):106–111, 2004.

[56] Timothy F. Brady, Viola S. Störmer, and George A. Alvarez. Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences*, 113(27):7459–7464, 2016.

[57] Timothy F. Brady, Viola S. Störmer, Anna Shafer-Skelton, Justin R. Williams, Andrew F. Chapman, and Heidi M. Schill. Scaling up visual attention and visual working memory to the real world. In *Psychology of Learning and Motivation*, volume 70, page 29–69. Academic Press, 2019.

[58] Brian J. Stankiewicz and Anuj A. Kalia. Acquisition of structural versus object landmark knowledge. *Journal of Experimental Psychology: Human Perception and*

Performance, 33(2):378–390, 2007.

[59] Mary Hegarty, Alan E. Richardson, Daniel R. Montello, K. Lovelace, and I. Subbiah. Development of a self-report measure of environmental spatial ability. *Intelligence*, 30(5):425–447, 2002.

[60] Maria Kozhevnikov and Mary Hegarty. A dissociation between object-manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, 29(5):745–756, 2001.

[61] Roy P. C. Kessels, Martine J. E. van Zandvoort, Albert Postma, L. Jaap Kappelle, and Edward H. F. de Haan. The corsi block-tapping task: Standardization and normative data. *Applied Neuropsychology*, 7(4):252–258, 2000.

[62] Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[63] Qi Wang, Daniel Smythe, Jun Cao, Zhilin Hu, Karl J. Proctor, Andrew P. Owens, and Yifan Zhao. Characterisation of cognitive load using machine learning classifiers of electroencephalogram data. *Sensors*, 23(20):8528, 2023.

[64] Farzana Khanam, A. B. M. Aowlad Hossain, and Mohiuddin Ahmad. Electroencephalogram-based cognitive load level classification using wavelet decomposition and support vector machine. *Brain-Computer Interfaces*, 10:1–15, 08 2022.

[65] Hong-Hai Nguyen, Ngumimi Karen Iyortsuun, Seungwon Kim, Hyung-Jeong Yang, and Soo-Hyung Kim. Mental workload estimation with electroencephalogram signals by combining multi-space deep models, 2024.

[66] Joshua K. Hartshorne and Laura T. Germine. When does cognitive functioning peak? the asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science*, 26(4):433–443, 2015.

[67] Ineke J. M. van der Ham, Michiel H. G. Claessen, Andrea W. M. Evers, and Melissa N. A. van der Kuil. Large-scale assessment of human navigation ability across the lifespan. *Scientific Reports*, 10(1):3299, 2020.

[68] Miloš Pušica, Aneta Kartali, Luka Bojović, Ivan Gligorijević, Jelena Jovanović, Maria Chiara Leva, and Bogdan Mijović. Mental workload classification and tasks detection in multitasking: Deep learning insights from eeg study. *Brain Sciences*, 14(2):149, 2024.

[69] Yingxin Liu, Yang Yu, Hong Tao, Zeqi Ye, Si Wang, Hao Li, Dewen Hu, Zongtan Zhou, and Ling-Li Zeng. Cognitive load prediction from multimodal physiological signals using multiview learning. *IEEE Journal of Biomedical and Health Informatics*, 29(5):3282–3292, 2025.