

PERCEIV: A Multimodal Physiological Dataset of Visual Encodings for Adaptive InfoVis Interfaces

Sebastian Idesis, Mireia Masias Bruns, Angela Lopez Cardona, Parvin Emami, Saravanakumar Duraisamy, Luis A. Leiva, and Ioannis Arapakis

Abstract—We contribute PERCEIV (PERception of visual Encodings in InfoVis), a large-scale dataset to support studies of user responses to visual encoding variants as a foundation for adaptive information visualization interfaces. The dataset includes recordings from 120 participants who completed tasks of varying difficulty, while simultaneous multimodal data were collected from brain, eye, and electrodermal activity. All data are time-synchronized and accompanied by relevant annotations and participant metadata. We release raw sensor streams together with fully processed event-aligned derivatives and rich trial-level annotations. We also provide baseline analyses and scripts for extracting cognitive load features across modalities. In addition, we make the code for data pre-processing, feature extraction, quality control, and training/evaluating baseline Machine Learning classifiers available. Standardized data splits and comprehensive documentation further enable reproducible benchmarking and rapid reuse. We hope our dataset will facilitate novel studies of sensor modalities, development of computational models, and deeper understanding of cognitive state dynamics in controlled information-processing contexts.

Index Terms—InfoVis, Cognitive Load, EEG, Eye gaze, EDA

I. INTRODUCTION

In today’s information-saturated world, users must decide what to process, filter, and prioritize. Information Visualisation (InfoVis) systems can address this challenge [1] and support critical decision-making by engaging suitable perceptual mechanisms [2]. In particular, physiological computing holds significant promise in InfoVis by incorporating human body signals to enable closed-loop, implicit monitoring and adaptive interfaces [3]. There is mounting evidence that cognitive processes are affected by interface properties and that such effects are detectable through physiological signals [4], [5], [6]. Cognitive load, for example, is measurable through changes in blood pressure [7], heart rate variability [8], brain activity [9], Electrodermal Activity (EDA) responses [10], and eye movements [9]. Building on this perspective, we consider user-centric, and context-aware solutions within a symbiosis-inspired Human–AI framework, where bidirectional communication taps into human cognitive resources to support decision-making [11].

Electroencephalography (EEG) is the most commonly used modality for assessing cognitive states [12], [13], [14], and prior studies have shown that EEG-derived indices like the

Task Load Index (TLI) can detect cognitive fit between task demands and visualisation formats [15], [16]. Despite this progress, most existing public datasets that offer EEG data focus on emotion recognition rather than on cognitive load analysis [17], [18], [19], with very few targeting visual perception [16], [20], thus highlighting a gap for InfoVis interfaces.

A. The Need for Multimodal Integration

EEG or Eye-tracking (ET) modalities alone only capture a subset of the user’s cognitive load. On the contrary, multimodal integration of these signals offers a more holistic view of ongoing cognitive dynamics [21]. This is particularly relevant for adaptive InfoVis interfaces, where different modalities respond to distinct task demands. For example, EEG reflects internal cognitive effort, whereas ET reveals visual attention and information foraging patterns. Within this broader context, the Cognitive Fit Theory [22] provides a foundation for understanding how alignment between visualisation formats and task types influences cognitive efficiency. Briefly, the theory proposes that problem solving is faster and more accurate when the information representation (e.g., tables, graphs) is in line with the cognitive operations required by the task, whereas any kind of dissonance may result in extra mental transformations that increase load and errors [22]. Building on this theory, our study introduces a new multimodal dataset designed to examine these interactions in real time using both physiological and behavioural indicators.

B. Contributions

Unlike many existing related datasets [23], [24], [25], which are often constrained by small sample sizes or limited modalities, our dataset PERCEIV¹ (PERception of visual Encodings in InfoVis) includes synchronized multimodal recordings from 120 non-expert participants, totalling 14,400 feature-rich trials, enabling more robust and generalisable analyses. Importantly, PERCEIV does not implement a closed-loop adaptive system. Instead, it records behavioural and physiological responses to controlled visual encoding variants, including predefined saliency-enhancing adaptations. These responses provide empirical evidence that can inform future user-model-driven adaptive InfoVis systems.

In this paper we provide a detailed description of PERCEIV, including descriptive information across behavioural and physiological measures, and explore how these measures relate

Sebastian Idesis, Mireia Masias Bruns, Angela Lopez Cardona, and Ioannis Arapakis are with Telefonica Scientific Research, Barcelona, Spain. Corresponding author: Sebastian Idesis (e-mail: sebastian.idesis@gmail.com).

Parvin Emami, Saravanakumar Duraisamy, and Luis A. Leiva are with the University of Luxembourg, Esch-sur-Alzette, Luxembourg.

¹<https://github.com/Telefonica-Scientific-Research/perceiv>

to task-induced, cognitive load features. In addition, we train Machine Learning (ML) models to predict said cognitive load features, with classification accuracy exceeding that of standard baselines, hence demonstrating our dataset's potential. Our main research question is the following: **How do diverse visualisation formats influence user performance, as measured by behavioural and neurophysiological signals, during problem-solving tasks?**

As an example application, we use our dataset to explore its potential for visual attention modelling in InfoVis tasks, a field of growing interest focused on understanding and predicting how users allocate attention when interacting with complex visual information [26]. Recent work has introduced the first models for goal-directed attention [27], [28] and, more specifically, question-driven saliency in InfoVis [2]. However, except very few studies (e.g., [29], [30], [31]), prior work has opted for crowdsourcing data collection methods over the more objective and reliable physiology-recording approaches. With this dataset, we take a first step in this research direction, aligning with the broader need for robust datasets that can also facilitate the adaptation of existing models to new visualization types and use cases [32].

II. RELATED WORK

Established theories of visualisation design consistently argue that visual representations should be optimized for rapid and intuitive interpretation [33], [34], [35], [36], [37], [38]. Previous research has examined the use of physiological signals in real-world systems. Take SmartHelm [39] as an example: a smart helmet for cargo cyclists that integrates EEG, ET and Augmented Reality interfaces to model the rider's attention and detect lapses in the wild. When reduced attention is detected, the system adapts the Augmented Reality assistance with cues and alerts, thereby closing the loop between multimodal physiological monitoring and adaptive feedback. Along this line, Toreini et al. [40] proposed attentive dashboards that rely solely on real-time ET data to deliver individualised attention feedback during visual exploration [40]. Finally, Karran et al. [41] used EEG-based neurofeedback in a business setting, applying an engagement index and a calibrated classifier to trigger adaptive interface changes, improving sustained attention and reducing error.

In InfoVis-specific work, Anderson et al. [16] used EEG to compare different box-plot variants, deriving trial-level cognitive load indices from spectral EEG features. Peck et al. [5] used fNIRS to evaluate bar graphs versus pie charts, relating prefrontal haemodynamic responses and NASA-TLX ratings of mental demand during visual judgment tasks. Both studies link visual design to cognitive workload using a single neuroimaging modality in controlled lab tasks with relatively small samples. We provide a larger-scale, multimodal dataset collected across multiple visualization types and complexity conditions, with released data and baselines to support reproducible modelling of task difficulty and cognitive demand in InfoVis settings.

Unlike previous studies that combine visual analysis tasks with explicit self-assessed workload labels, our dataset pro-

vides synchronized multimodal recordings from novice participants under systematically manipulated task conditions. Rather than framing these manipulations as direct measures of cognitive load, we treat them as controlled experimental factors that modulate expected cognitive demands.

A. Approaches to Multimodal Mental Load Assessment

Traditionally, behavioural measures have been used to evaluate mental load and cognitive states. However, near-identical behavioural performance, such as Accuracy or Reaction Times (RTs), may mask differences in underlying cognitive effort [42], suggesting the need for more ecologically-valid physiological assessments. Moreover, many behavioural measures (e.g., task completion time) are sparse and event-bound, observable only at response or trial completion, thereby yielding limited information during ongoing processing. This sparsity complicates implicit, closed-loop modelling that requires continuous, time-resolved estimates of cognitive state for enabling dynamic adaptation. For example, [15] applied a neuroergonomics framework to investigate cognitive fit using EEG and found that graphical visualisations reduced both subjective and physiological-derived workload metrics, when aligned with spatial task demands. Their results validated the EEG-derived TLI as a reliable workload metric in InfoVis settings.

Recent studies have focused on multimodal integration. For example, Haapalainen et al. [43] used elementary cognitive tasks to induce varying levels of mental engagement while collecting ET, heart activity, EEG, and skin conductance data. Beyond feasibility, multimodal fusion appears to enhance robustness by exploiting complementary sensitivities, mitigating modality-specific noise or dropouts, and enabling earlier, more reliable detection of workload shifts than any single signal alone. Along this line, Zuger et al. [44] further demonstrated that features such as EEG patterns, eye blinks, EDA, heart rate, and inter-beat intervals correlate with user interruptibility, a factor that is negatively associated with cognitive mental load. These studies collectively highlight the value of combining central (e.g., EEG) and peripheral (e.g., EDA) physiological signals to improve mental state estimation [45].

B. Current Datasets on Multimodal Cognitive Workload

Due to their intuitive operation, utility, and relatively low cost, eye trackers have been favoured in InfoVis [46], [47], [28], [2], [29], as they offer valuable insights into information foraging and decision-making processes, thereby enabling the curation of many public datasets that leverage exclusively this modality. The most popular dataset for visual perception analysis of InfoVis interfaces is MASSVIS [46], which consists of eye movements by 33 participants, on 393 visualisations. Platsek et al. [48] leveraged MASSVIS to explore how user attention and perception influence task-based analysis. Wang et al. [2] created the SalChartQA dataset, using the BubbleView technique [49], to supplement the use of an actual eye-tracker, and demonstrated a strong effect of including questions on visual saliency prediction. Chang et al. [26] presented a new dataset of saliency maps along with three literacy assessment

TABLE I: Overview of existing multimodal cognitive workload datasets, sorted from lower to higher amount of participants.

Dataset & Reference	Modalities	N	Task Type	Task Difficulty Manipulation
CogLoad [50]	EDA TEMP HR ACC	23	Six PC-based cognitive tasks (easy/med/hard)	Task difficulty + secondary task + NASA-TLX
CLARE [51]	EEG EDA ET ECG	24	MATB-II multitasking	Continuous cognitive load ratings every 10s
Kalatzis et al. [52]	ECG RR	26	Tracking task + Complementary and simultaneous task	Task complexity + NASA-TLX
ADABase [53]	ECG EDA ET EMG PPG Behavior	51	n-back and k-drive test	Task complexity + NASA-TLX
CLAS [54]	EEG PPG ECG ACC	62	Math problems, Logic problems and Stroop test	Task difficulty; Time pressure
Ours	EEG EDA ET PPG Behavior	120	Navigation of distinct visualisation types	Question difficulty, stimuli complexity, and visual adaptations

scores, also using the BubbleView technique, to provide insights into the effect of individual differences in saliency modelling. Unfortunately, none of the aforementioned InfoVis-dedicated datasets considers multimodal physiological data.

Multimodal datasets have been developed in other domains for cognitive load assessment, sparking new research developments and interpretations. Table I provides a summary of related publicly available datasets that incorporate physiological sensors like EEG, Blood Volume Pulse (BVP), EDA, respiration rate, and ET.

Specifically, Gjoreski et al. [50] created two datasets (Snake and CogLoad) in which 46 participants performed smartphone and PC-based tasks of varying difficulty, while heart rate, EDA, skin temperature, and acceleration were recorded via wrist-worn devices. Their work validated the feasibility of predicting cognitive load using physiological signals (in the range of 68-82% accuracy) using SVM classifiers. These datasets do not include EEG, as their focus was on wearable sensing rather than high-resolution neurophysiological recordings.

The CLARE dataset [51] consists of EEG, electrocardiogram (ECG), EDA, and gaze data from 24 participants. The dataset includes real-time self-reports of cognitive load every ten seconds, enabling fine-grained labelling for supervised machine learning approaches. In addition, the CLAS dataset [54] contains multimodal physiological recordings from 62 participants during a cognitively demanding task-switching paradigm that involves arithmetic and memory tasks. Furthermore, Kalatzis et al. [52] presented a multimodal dataset for classifying cognitive workload using ECG and respiration signals. The data were collected from 26 participants across low and high workload conditions and it was validated using NASA-TLX, HRV, and RRV metrics. Finally, the ADABase dataset [53] provides multimodal recordings (EEG, ECG, EDA, respiration, and pupil data) from 53 participants performing a symbol-digit matching task under varying time constraints, enabling fine-grained cognitive load estimation using physiological and behavioural signals.

C. Limitations of Current Datasets

Most datasets rely on small sample sizes or lack proper synchronization across modalities. While some provide multimodal data, they often exclude high-resolution neural recordings such as EEG, omit detailed behavioural annotations, or fail to incorporate adaptive task elements. Addressing the aforementioned limitations, the PERCEIV dataset specifically targets InfoVis tasks with a focused, practically deployable set of modalities. Specifically, we combine EEG (cognitive effort), ET (visual attention), EDA (autonomic arousal), and behavioural measures (accuracy, RT), yielding complementary views on user state during navigation of complex visualisations. Unlike prior work that either uses a single modality in InfoVis or multimodal sensing in non-visualisation tasks, PERCEIV situates this multimodal collection directly within realistic InfoVis scenarios and across multiple visualisation types.

III. METHODS

Data were collected at two institutions, hereafter referred to as Site 1 and Site 2. The study protocol was reviewed and approved by the Ethical Review Board of [redacted] with ID [redacted].

A. Stimuli

All visualisations in our study show information about a criminal network and the interaction between its members. A schematic illustration of each visualization type is presented in Figure 1 while a real example of each type is shown in the Supplementary Materials. Specifically:

- **Graphs:** Nodes in the graph represent individuals, while the edges that connect them indicate an association. Graphs also contained information about people's name, gender, number of arrests, whether they are in prison, their birthplace, and the connection strength with other individuals [55], [32].
- **Timelines:** Timeline visualisations reveal the history of phone calls between individuals at different times of the day, indicating the person who is calling (caller), the

person that receives the call (callee), and the duration of the call.

- **Distribution charts:** Distribution chart visualisations indicate the number of calls at a set granularity level; in our case at each hour of the day.

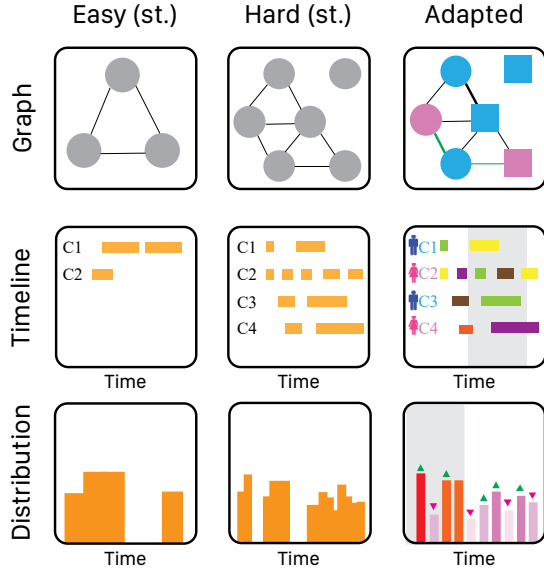


Fig. 1: Schematic representation of visualisation types, with an adaptive version of the high complexity visualizations.

B. Participants

A total of 120 participants (75 male, 45 female) aged 28 (SD=7.5) years took part in the study. Participants were equally distributed across institutions (N=60), using the same equipment except the EEG device. For each visualisation type, 20 participants were randomly assigned, resulting in 40 participants per visualisation type (2 sites $\times$ 3 types of visualisation $\times$ 20 participants). Across visualization types and institutions, age distributions were comparable (means ranging from 24.8 to 29.5 years; SDs 3.7–9.6), and gender ratios followed a similar pattern across sites (e.g., Graph: 16/4 and 11/9 male/female; Distribution: 11/9 and 13/7; Timeline: 12/8 and 12/8). Detailed demographics of each participant group are provided in the Supplementary Materials.

Participants were recruited through an online testers platform. All participants provided written informed consent prior to participation and received monetary compensation (\$30 for 1 h of study on average). Regarding the exclusion criteria, a few more participants were initially recruited, but only those with data of acceptable quality (as defined by our preprocessing procedures) were retained in the final dataset sample. All participants had normal or corrected-to-normal vision and none of the participants reported a history of neurological or psychiatric disorders, nor were they taking any medication that could interfere with the results.

C. Procedure

The experiment began with a warm-up task, followed by a set of eight practice trials, to ensure that participants understood the protocol. Each session consisted of 120 trials per participant. Each trial implemented the following steps (see Figure 2B):

- **Fixation cross:** At the centre of the screen, for 1 s.
- **Question screen:** Question about the upcoming visualisation. Self-paced response guarantees that the user has enough time to read the question.
- **Visualisation screen:** Until the subject responds, with a timeout of 10 s. Trials reaching the 10 s timeout were retained, assigned a reaction time of 10 s, and labelled as incorrect responses. The possible answers are single-digit numbers in the $[0, 9]$ range.
- **Resting period:** Self-paced response screen to rest.

All tasks were presented on a 1920×1080 px monitor. The presentation order of the stimuli was randomised, to mitigate learning or carry-over effects. Further, we examined accuracy and reaction time as a function of trial index and no systematic trend toward improved accuracy or faster responses was observed across trials (see Supplementary Material), suggesting that no learning effects were present.

D. Experimental Design

1) **Independent Variables:** The experiment was designed to systematically manipulate task difficulty through two independent variables: (i) question difficulty and (ii) stimulus complexity (Figure 2A). Each variable was defined at two levels (low/high) and implemented orthogonally across trials, to allow for factorial analysis of their effects on user performance and physiological responses. Each participant was randomly assigned to only one visualisation type (between-groups design), to prevent potential learning or exposure effects across conditions. Consequently, comparisons across visualization families should be interpreted cautiously, as visualization type was not manipulated within participant.

Question difficulty was controlled by adjusting the semantic and numerical reasoning demands required to answer the questions. Low-difficulty questions involved simple feature detection or direct recall, while high-difficulty questions required integrating multiple elements or inferring indirect relationships within the visualisation. Specifically, an easy question queried a single visualization attribute (e.g., “How many females are shown?”), whereas a hard question required combining two attributes (e.g., “How many females have been arrested more than 10 times?”).

Stimulus complexity was controlled by modifying the amount and density of visual information displayed. In the low-complexity condition, visualisations included fewer nodes, edges, or events (depending on the visualisation type), while the high-complexity condition introduced a larger number of elements and more dense structures. For graphs, low-complexity stimuli contained 3 nodes, whereas high-complexity stimuli contained 6 nodes. For timelines, low-complexity stimuli showed 2 callers over 5 hours, while high-complexity timelines displayed 4 callers over 12 hours. For

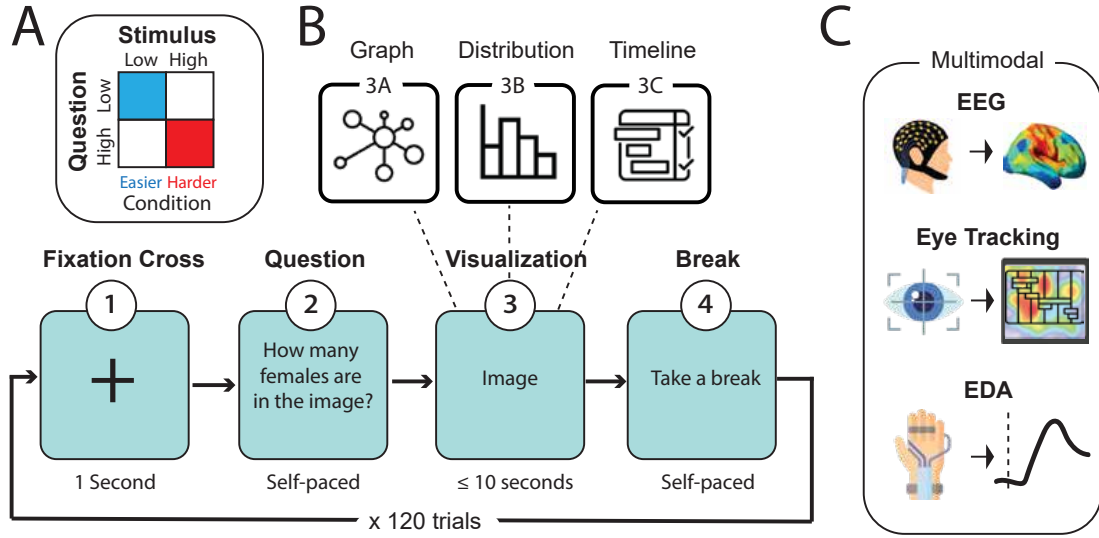


Fig. 2: (A) Task design: A 2×2 factorial grid crosses Question difficulty (low/high) with Stimulus complexity (low/high); “easier” corresponds to low–low and “harder” to high–high combinations. (B) Trial timeline: (1) fixation (for 1 s), (2) question (self-paced), (3) visualisation and response (< 10 s) from one of three families (Graph, Distribution, or Timeline) followed by (4) a self-paced break; each participant completed 120 trials. (C) Synchronized multimodal acquisition records: EEG, ET, and EDA, all time-aligned with behavioral events.

distributions, low-complexity plots covered 5 hours of activity, and high-complexity plots covered 24 hours.

For a fourth of the trials (40 in total), combined visual adaptations were introduced with the aim to support faster cognitive processing. There, the visualisations changed properties such as node size, color encoding, or edge thickness to emphasise task-relevant information. These adaptations were not triggered dynamically based on user performance or behavioural traits; instead, they were predefined modifications to the visual encoding, designed to make elements more visually distinct. Thus, in our dataset, “adaptation” refers to visual encoding variants rather than real-time interface adaptations. The goal was to capture how users respond to encoding changes that could later be selected dynamically by an adaptive pipeline, such as the one presented by Lopez-Cardona et al. [56].

These 3 factors (question difficulty, stimulus complexity, stimulus adaptation) were pseudo-randomly assigned across trials, ensuring a balanced representation ($2 \times 2 \times 2$) across participants. Specifically, the 120 trials comprised adapted and non-adapted trials, with both adaptation levels equally distributed across the four question-difficulty $\times$ stimulus-complexity conditions. Therefore, adaptation effects were assessed independently of question difficulty and stimulus complexity. This design enabled the systematic investigation of how different visualisation strategies impact user performance, as measured by behavioural performance and multimodal physiological signals (see Figure 2C).

Task difficulty and stimulus complexity were used as experimental manipulations intended to induce variations in cognitive demand. However, we do not treat these factors as ground-truth labels of cognitive load. Individual differences could lead to high cognitive effort even on nominally “easy” trials, or low effort on “hard” ones. PERCEIV therefore cap-

tures behavioural and physiological responses under controlled difficulty conditions, without assuming that these conditions directly map onto absolute cognitive load levels.

2) *Behavioural Measures*: To evaluate user performance and engagement across conditions, we collected behavioural measures throughout the task. These included response accuracy (correct vs. incorrect answers to task-specific questions) and RT (measured from stimulus onset to user response). These behavioural indicators were used to assess how different visualisation types and adaptation strategies impacted task efficiency and cognitive demand.

3) *Timing and Synchronization Strategy*: The experimental task was implemented using PsychoPy [57], a widely used software in cognitive and behavioural experiments. It provides millisecond-level timing accuracy and supports precise synchronization with external physiological devices. Specifically, PsychoPy presented the visualisations and recorded participants’ answers. To ensure synchronized data collection, all devices were time-aligned using a shared timestamp protocol [58]. They operated on the same local network, maintaining consistent internal clocks and allowing for precise temporal alignment across all data streams.

E. Sensory Data Acquisition and Preprocessing

1) *EDA*: EDA measures skin conductance levels, capturing variations in physiological arousal. The EDA signal was processed differently, depending on the site. At the first site, the EEG amplifier device included two electrodes that were connected to the participant fingers (distal phalanges of the index /middle fingers of participants’ non-dominant hand), providing data from the EDA and therefore, the signal was encoded as an extra channel with a sampling rate of 250 Hz.

At the second site, the EDA signal was recorded with a ring-type sensor, also in the participants' non-dominant hand. For this site, we used a stand-alone device (Bitbrain Ring) that operated at a sampling rate of 32 Hz. PPG was also recorded at both sites, using the same sampling rate as the corresponding EDA signal at each site (250 Hz at Site 1 and 32 Hz at Site 2).

2) *EEG*: Site 1 used a 32-channel, gel-based actiCHamp EEG system (Brain Products GmbH, Germany). A high-impedance amplifier ensured optimal Signal-to-Noise Ratio (SNR). Electrode placement followed the international 10–20 system. The ground and reference electrodes were positioned at Fpz and Cz, respectively. Impedances were kept below 10 k Ω , following standard procedures (see the Supplementary material for additional information). Electrode preparation took approximately 15 minutes per participant. Before each recording, the scalp was cleaned and mildly abraded to reduce impedance. Visual inspection ensured all electrodes were properly attached and impedance levels remained within acceptable thresholds throughout the session. Additionally, EEG channels contaminated with persistent noise were interpolated using spherical splines, when necessary. This was in accordance to previous work (e.g., [59]) that used said device to conduct a quality assessment study.

Site 2 used a 12-channel, dry-based EEG device (Bitbrain Diadem system). Electrodes A1 and Fpz were used for reference and ground, respectively, following the international 10–10 system. Raw EEG signals were recorded for subsequent processing. Setup time was approximately five minutes and signal quality was monitored via the system's internal quality indicators (see the Supplementary Materials for additional information). Electrode contact was verified using the system's built-in quality indicators and no skin abrasion or gel was required. Preprocessing steps mirrored those applied to the gel-based data. Channels flagged as noisy or unstable were excluded from analysis or interpolated where appropriate. Previous studies (e.g., [60]) have used the aforementioned device to conduct a quality assessment study.

For both devices, signals were sampled at 256 Hz, offering sufficient temporal resolution for downstream analysis. Artifact correction was carried out using Independent Component Analysis (ICA), thus eliminating ocular and muscle-related components. Furthermore, recordings took place in sound-attenuated, electrically shielded rooms to minimize external interference.

Because the EEG systems differed in electrode type and channel density, we treat site as an explicit dataset factor rather than assuming that both EEG subsets are fully interchangeable. All pooled EEG analyses reported below are therefore based on site-normalized features and are accompanied by site-specific analyses

3) *Eye Tracking*: Eye movements were collected at both sites using the GP3 HD Eye Tracker, sampling at 150 Hz in accordance with recommended guidelines [61], [47], [62]. This equipment provides high spatial and temporal resolution [63], [64]. A five-point calibration was conducted for each participant before the study, and it was repeated until minimal error was achieved. Throughout the experiment, various eye

movements were recorded, including fixations, saccades, and blinks. The data collected included timestamps, gaze coordinates (x,y), and eye validity codes, with continuous monitoring of data quality through real-time feedback [65], [66].

4) *Feature extraction*: Regarding the EEG data, offline preprocessing was performed in Python and MATLAB, including 50 Hz notch filtering, band-pass filtering between 0.5–40 Hz, artifact rejection, and epoching. Baseline correction was applied using the 500 ms pre-stimulus interval. We then computed Welch's Power Spectral Densities (PSDs) and band-limited measures over the canonical bands [67], [68]: δ (0.5–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–20 Hz), γ (30–40 Hz).

The EEG features used in our analysis included:

- **Frontal Theta power**: integral of the mean PSD restricted to the frontal Regions of Interest (ROIs).
- **Frontal Alpha Asymmetry (FAA)**: difference of alpha power between the electrodes F3 and F4.
- **Spectral and amplitude entropy**: Spectral and amplitude entropy of the channel-averaged time-series amplitude distribution.
- **Theta Phase Locking Value (PLV)**: phase-locking values across channel pairs in the theta frequency band.

The selected EEG features were chosen to cover complementary dimensions of task-related neural activity previously associated with cognitive effort and working-memory demands: spectral power, hemispheric asymmetry, signal irregularity, and functional connectivity [69]

Regarding eye data, we computed the following features:

- **Pupil dilation**: Relative change of pupil size after the appearance of the relevant stimuli
- **Fixation amount**: Number of fixations
- **Fixation duration**: Mean fixation duration (in seconds)
- **Saliency coverage**: Amount of space captured by the saliency in the whole screen [70].
- **Saliency entropy**: Entropy of the spatial saliency distribution, quantifying how dispersed or concentrated visual attention is across the stimulus. [64].

Finally, we computed the following features from EDA data:

- **Area under the curve**: Simpson integral of the baseline-corrected EDA signal over the trial window.
- **Number of peaks**: Count of local maxima with height higher than mean plus one standard deviation.
- **Time to peak**: Latency to the first detected peak (in seconds).
- **Recovery time**: Time from peak to 50% amplitude decay (half-recovery).
- **Peak amplitude**: Amplitude of the first detected peak.

An extended explanation of all these features is provided in the Supplementary Materials. Although PPG recordings are included in the released dataset, they were not included in the baseline feature analyses because their overall signal quality was insufficient for reliable feature extraction across participants.

IV. RESULTS

A. Condition Differentiability Analysis

To assess potential device confounds, we repeated the condition analyses with Site as a fixed factor, using mixed-effects models. For behavioural measures, the main Easy–Hard effects remained robust for accuracy and RT, whereas site effects were small and not statistically reliable. The same pattern was observed for EDA and ET after within-site normalization. For EEG, features were first z-scored within site, and mixed-effects models showed that the main Easy–Hard effects remained robust, whereas Site main effects were small and not statistically reliable, with no meaningful Site $\times$ Condition interactions. Site-specific analyses, including the magnitude of Site effects and Site $\times$ Condition interactions for each modality, are reported in Supplementary Section J. Given the manuscript page-limit constrain, classifier performance is retained as a pooled, site-normalized baseline rather than expanded into an additional per-site figure. A dedicated study using the same dataset provides a more extensive comparison of the two EEG systems across signal-quality and analytical metrics [71]. Because visualization family was assigned between participants, the analyses below focus primarily on Easy–Hard effects within each visualization group. Therefore, unless explicitly stated otherwise, all reported effect-size ranges summarize separate Easy–Hard comparisons conducted within each visualization group and should not be interpreted as direct statistical comparisons between visualization families. As a summary, differences in effect-size magnitude across graph, timeline, and distribution conditions are reported descriptively and should not be interpreted as direct within-subject comparisons between visualization families.

1) *Behaviour*: We examined participants’ accuracy and RT across conditions of task complexity and cognitive load features, for three distinct types of visualisations: (i) graphs, (ii) timelines, and (iii) distribution charts (Figure 3).

In terms of accuracy, a significant main effect of task difficulty was consistently observed within each visualisation group. Within each visualisation group, participants were more accurate in the Easy conditions (low stimuli load) compared to the Hard conditions (high stimuli load) ($p < .01$ Cohen’s d range across groups: 1.102–4.28). Additionally, accuracy declined significantly with increasing question complexity within each visualisation group ($p < .01$, Cohen’s d range across groups: 0.55–3.2). Adaptation played a major role in mitigating this effect. Fully adapted visualisations resulted in higher accuracy scores ($p < .01$, Cohen’s d between 0.56 and 1.42 across visualisation types) than baseline (non-adaptive) visualisations (Figure 3A).

RT analyses revealed a complementary pattern within each visualisation group. Participants responded significantly faster in the Easy questions condition compared to the Hard questions condition ($p < .01$ Cohen’s d between -1.571 and -4.591 across visualisation types), regardless of the visualisation type. RT increased as stimulus complexity rose ($p < .01$ Cohen’s d between -3.541 and -6.426 across visualisation types). Importantly, adapted encoding led to significantly reduced RT compared to baseline ($p < .01$ Cohen’s d between -

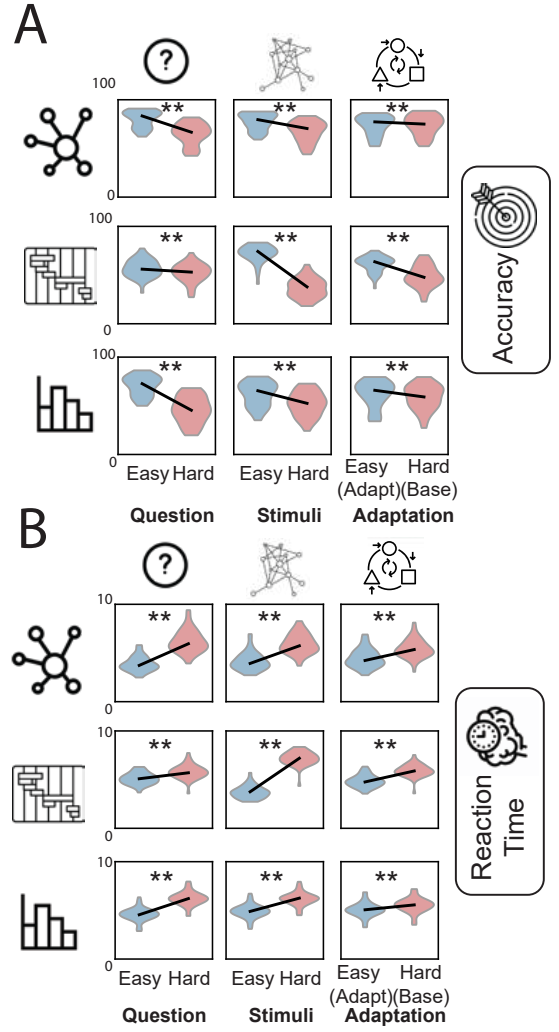


Fig. 3: Behavioural results across visualisation types: (A) Accuracy (%) for Easy vs. Hard conditions across the three comparisons: question difficulty, stimulus complexity, and adaptation level. For the adaptation comparison, “Hard (Base)” and “Easy (Adapt)” are interpretive labels indicating the expected relative difficulty of the non-adapted and adapted encodings, respectively. Adapted and non-adapted trials were equally distributed across all four question-difficulty $\times$ stimulus-complexity conditions, so the adaptation contrast was assessed independently of those two factors. Rows correspond to the three visualisation types (graph, timeline, distribution). (B) Reaction time (s) for the same comparisons and visualisation types. Violin plots show participant-level distributions, and black lines connect medians. Asterisks denote significant differences ($p < .05$: *, $p < .01$: **, FDR corrected).

1.068 and -2.985 across visualisation types), suggesting a cognitive facilitation effect afforded by adaptive visualisation (Figure 3B).

Because EEG was acquired with different systems at the two sites, we first examined site/device effects using mixed-effects models with Site as a fixed factor. The main Easy–Hard effects were directionally consistent across sites after within-

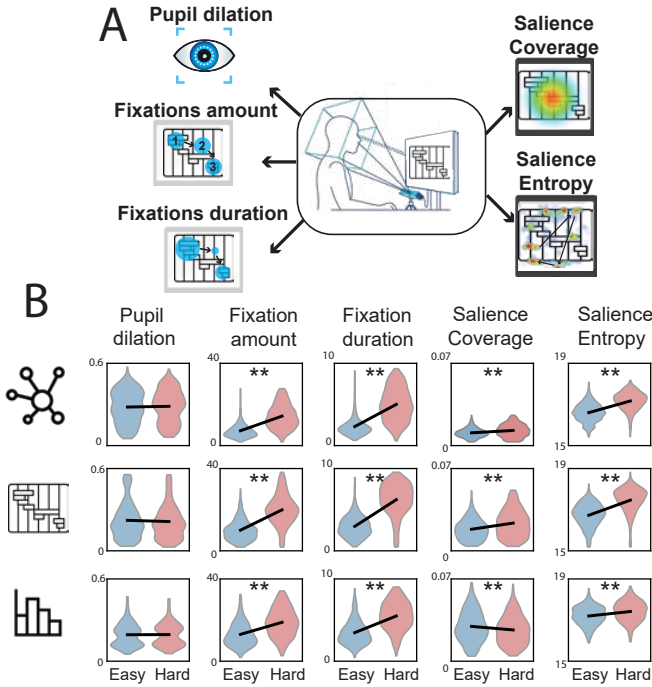


Fig. 4: Eye-tracking pipeline and workload effects. (A) Experimental setup and trial-level features extracted from the gaze stream. (B) Violin plots comparing Easy (blue) vs Hard (red) conditions for the three visualisation types (Graph, Timeline, Distribution: icons at left). Black horizontal bars indicate medians across subjects. Asterisks mark significant Hard–Easy differences in the paired t-test across subjects ($p < .05$: *, $p < .01$: **, FDR corrected). Pupil dilation did not survive FDR correction.

site normalization (See Supplementary Material). We therefore report the site-normalized aggregate EEG results below as a summary, while the full Site-specific EEG tables (N=60 per site) are provided in the Supp. Materials to make the hardware dependence transparent. This interpretation is also supported by recent work comparing dry- and gel-based EEG quality metrics in InfoVis environments [72].

2) *ET*: We evaluated three gaze-based features: (i) baseline-corrected pupil dilation, (ii) fixation amount (number of fixations), and (iii) fixation duration. Furthermore, we inspected two saliency-based metrics, namely (i) saliency coverage, and (ii) saliency entropy (Figure 4A).

We averaged trials within each condition (Easy: low complexity/low stimuli load, Hard: high question complexity/high stimuli load) for each participant and visualisation, and then ran paired within-subject tests (Hard vs. Easy, FDR-corrected across comparisons). In what follows, we report the results from participants from both acquisition sites (N=120) and reserve separated results for each site of acquisition (N=60 per site) in the Supplementary Materials.

Pupil dilation was generally greater in Hard conditions compared to Easy conditions across the three visualization types. Although this aligns with classic workload effects, the differences were modest (Figure 4B, left column) and were

not statistically significant after applying correction ($p > .05$, Cohen’s d between -0.28 and -0.37 across visualisation types, FDR-corrected). To avoid ambiguity, the pupil panels in Figure 4B are not marked with significance asterisks. In contrast, fixation amount increased in Hard conditions, indicating a more pronounced visual search.

This effect was observed within each visualisation group (Figure 4B; in all groups, Hard > Easy, $p < .01$, Cohen’s d range across groups: 2.21–3.38, FDR-corrected) (Figure 4B, middle-left column). Similarly, fixation duration, saliency coverage, and saliency entropy were all significantly higher in the Hard condition within each visualisation group (Hard > Easy, $p < .01$ Cohen’s d between, FD: 2.97 and 4.19, SC: 0.97 and 1.68, SE:1.26 and 3.52, across visualisation types, FDR-corrected).

Taken together, the ET results demonstrate that, within each visualisation group, higher task difficulty primarily manifests as more frequent and longer fixations (i.e., more exploratory sampling of the visualisation) and higher saliency coverage/entropy, while pupil-based effects are more variable.

3) *EEG*: We observed several EEG features associated with increased task difficulty, although the effects were not uniform across all visualization groups (Figure 5A). Because EEG was acquired with different systems, we tested the device effects before interpreting aggregate EEG results. EEG features were first z-scored within site, and mixed-effects models were then fitted with Condition and Site as fixed factors. Across the EEG features, the main Easy–Hard effects remained robust, whereas Site main effects were small and not statistically reliable. This suggests that the workload-related EEG effects were broadly consistent across sites after normalization. Therefore, below, we report the results from participants from both acquisition sites (N=120) and reserve separated results for each site of acquisition (N=60 per site) as Supp. Materials.

Specifically, theta-band power increased in the Hard condition, especially when restricted to the frontal ROIs. These effects held for graph ($p < .01$, Cohen’s $d = 0.94$) and timeline ($p < .05$, Cohen’s $d = 0.77$) visualisations but were not confirmed during multiple-comparisons control ($p > .05$, Cohen’s $d = 0.30$) in distribution visualisation (see Figure 5B left). Similarly, FAA (difference between F3 and F4 power) increased in the Hard condition, an effect that is consistent with heightened cognitive/affective demand. This FAA increase was significant for two visualisations ($p < .05$, Cohen’s $d = 0.55$ and $p < .01$, Cohen’s $d = 0.67$) and showed the same tendency for the distribution visualisation ($p > .05$, Cohen’s $d = 0.22$), see Figure 5B center/left).

Measures of signal irregularity revealed expected patterns: Amplitude entropy was higher under Hard conditions (more irregular and less structured activity), with statistically significant differences ($p < .01$, Cohen’s d between 1.28 and 1.32 across visualisation types) after correction in all three visualisations (see Figure 5B centre/right). Functional connectivity decreased with difficulty: theta-band PLV was lower for Hard than Easy trials (significant for graph visualisations, $p < .05$ Cohen’s $d -0.97$), showing no significant difference in the remaining two, see Figure 5B right). To make the site/device dependence transparent, we also report EEG results separately

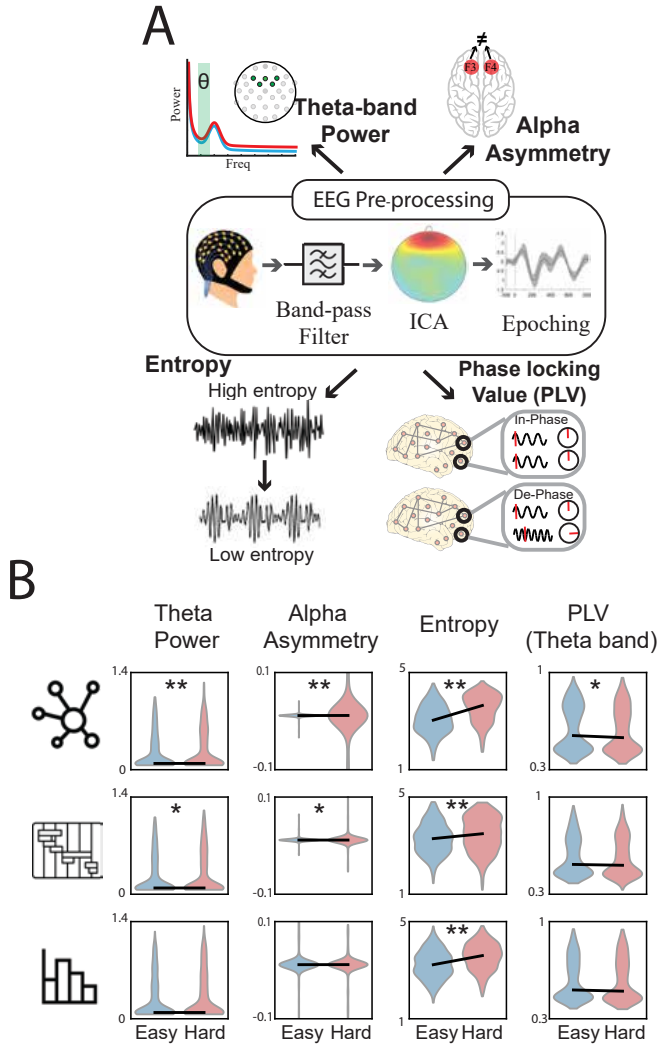


Fig. 5: EEG workload metrics. (A) Processing pipeline: continuous EEG is band-pass filtered (0.5–40 Hz), cleaned with ICA to remove ocular/muscle artifacts, and epoched (0.5 to 3 s). From the post-stimulus window we extract the analyzed metrics (B) Within-subject comparisons of Easy (blue) vs Hard (red) for each visualisation type (rows: Graph, Distribution, Timeline). Violin plots display participant means: black lines connect medians across conditions. Asterisks denote paired t-tests (Hard vs. Easy, $p < .05$: *, $p < .01$: **, FDR corrected).

for both sites in the Supplementary Material, including all EEG features used in the main analyses.

4) *EDA*: To assess peripheral physiological responses in response to task complexity, we analysed EDA signals and extracted five time-domain features per trial: (i) Area Under the Curve (AUC), (ii) number of peaks, (iii) time to peak, (iv) recovery time, and (v) peak amplitude (Figure 6A).

Each metric was averaged per subject and compared between Easy and Hard trials, using paired t -tests. To account for multiple comparisons, all p -values were FDR-corrected using the Benjamini-Hochberg method. Here, we report the results from participants from both acquisition sites ($N=120$)

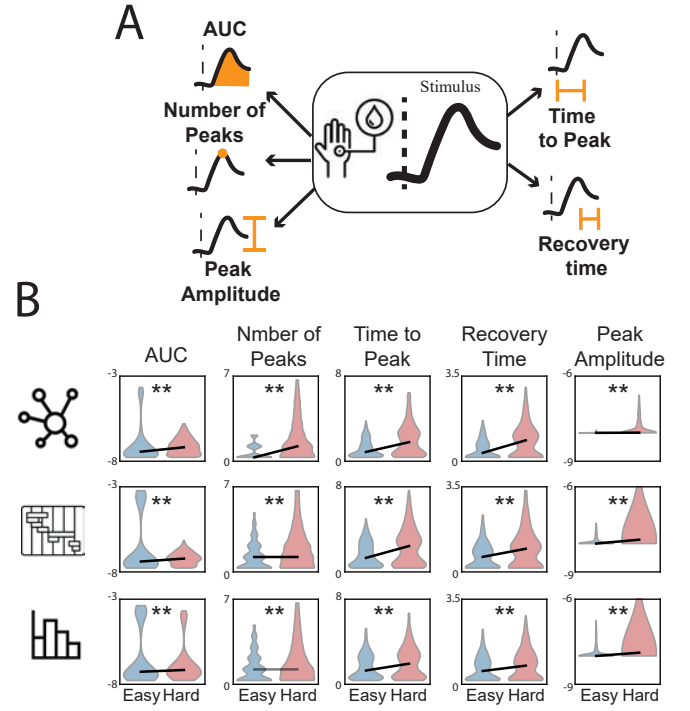


Fig. 6: EDA signal features and effects of task difficulty. (A) Schematic illustration of the five EDA features extracted. (B) Violin plots of each metric comparing Easy and Hard conditions for each visualisation type (Graph, Timeline, Distribution). Lines connect paired median values. Asterisks indicate significant differences ($p < .05$: *, $p < .01$: **, FDR corrected). All metrics show increases under high difficulty, especially for peak-based dynamics.

and reserve separated results for each site of acquisition ($N=60$ per site) as Supplementary Materials.

In the Graph condition, all EDA features showed a statistically significant increase ($p < .01$ Cohen's d between 0.27 and 2.33 across visualisation types) under high workload, after FDR correction (Figure 6B). The same Easy–Hard pattern was observed within the Timeline and Distribution groups across all EDA metrics ($p < .01$; Cohen's d range: Timeline = 0.37–0.91, Distribution = 0.27–0.67). These effect-size ranges are reported descriptively and are not intended as direct statistical comparisons between visualization families. Overall, the results presented in Figure 6 confirm that EDA dynamics, especially peak-based features, are sensitive to task-induced cognitive differences within each visualization group. These findings suggest that EDA metrics can effectively estimate features for cognitive load-centric InfoVis adaptations.

B. Cross-Modality Alignment

We first trained linear SVMs (subject-grouped, stratified 5-fold CV) to assess modality-specific discriminability (see Figure 7A), before extending the analysis to a broader suite of classifiers (Figure 8). Overall, behavioural features were the most informative: RT-only outperformed Accuracy-only, while

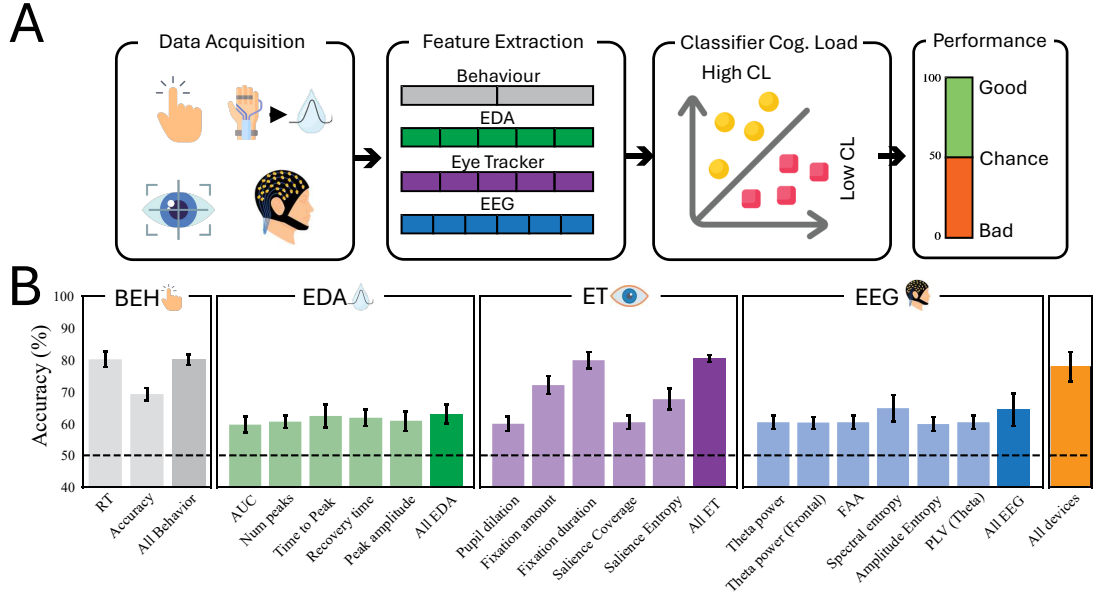


Fig. 7: Multimodal workload classification: (A) Pipeline overview. Per-trial features are extracted from each modality and standardized prior to input to binary classifiers that predict Easy (low workload) vs Hard (high workload). Models are evaluated with subject-grouped cross-validation to avoid leakage (stratified by class and grouped by participant). (B) Example performance of the linear SVM for each feature set. Error bars denote standard deviation. The rightmost panel reports performance when all modalities are used together. The dashed line marks chance level performance.

RT+Accuracy yielded the best performance (0.80–0.85 mean Accuracy).

Physiological modalities were less discriminative, but consistently above chance. For EDA, individual features achieved mid-50s to mid-60s accuracy, with peak-amplitude and time-to-peak among the strongest, whereas combining all EDA features improved performance into the high-60s/70s percent range. EEG features appeared to have more discriminatory power, with spectral entropy performing best. Combining all EEG features reached 0.60–0.66 Accuracy. Lastly, ET signals were the most informative within the physiological devices, achieving an Accuracy of about 0.80 (see Figure 7B). These results indicate that ET carried the strongest modality-specific physiological signal in the present benchmark, whereas EEG showed lower classification performance and greater sensitivity to differences between acquisition systems.

Notably, combining all physiological modalities (i.e. EDA+ET+EEG) improved the performance of the classifiers over any single physiological set (0.75–0.85), making them comparable to the behaviour-only models. The discriminatory power of the ET features validate the popularity of this family of signals and suggest the importance of generating further synthetic datasets for larger models.

Because the Easy/Hard contrast combines question difficulty and stimulus complexity, we additionally evaluated two auxiliary classification tasks based on the orthogonal 2×2 design: low vs. high question difficulty and low vs. high stimulus complexity. These auxiliary analyses use the same

subject-grouped cross-validation pipeline and are reported in the Supplementary Material.

It is relevant to clarify that for EEG-based and multimodal classifiers, site was treated as a potential source of heterogeneity. EEG features were normalized within site before modeling, and cross-validation was grouped by participant to prevent subject-level leakage. Within each cross-validation fold, the site-specific normalization parameters were estimated exclusively from the training participants and then applied to the held-out participants from the corresponding site. We interpret pooled EEG classification results as aggregate baselines over the released dataset rather than as evidence that the recordings of both EEG devices are interchangeable.

Figure 8 shows performance across all classifiers, including linear and quadratic discriminant analysis [73], logistic regression [74], naïve Bayes, Polynomial SVM [75], random forests [76], gradient boosting [77], and k-nearest neighbours [78]. We use nested, subject-independent cross-validation to estimate generalisation to unseen participants. In each outer fold (5 folds), a small set of subjects was held out for testing, and all data from those subjects were excluded from training (leave-subjects-out, no subject overlap). Within each outer training set, features were ranked by mutual information, candidate top-K were generated, and a grid search over hyperparameters and feature subsets was run using inner subject-wise cross-validation (3 folds) (See Supplementary Materials for results with optimal features). Preprocessing steps, including median imputation and standardisation, were fitted exclusively on the training folds. Models selected in the

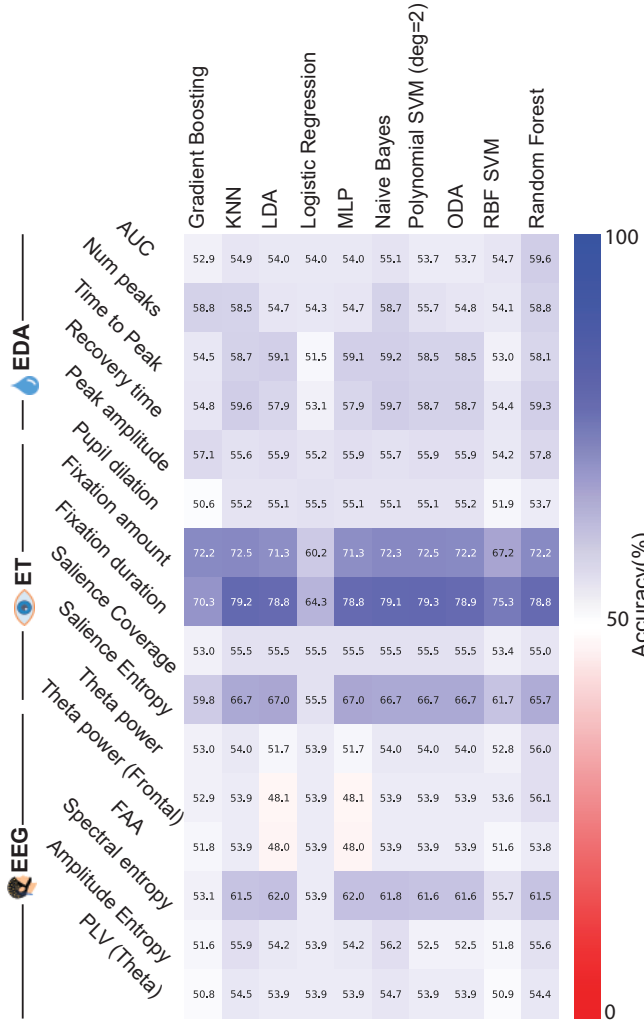


Fig. 8: Accuracy of Multiple classifiers: Features are standardized and fed to binary classifiers that predict Easy (low workload) vs Hard (high workload). Accuracy is reported for each metric for each different classifier. Values above chance level (50 percent) are depicted in blue and values lower than chance level are depicted in red, revealing how almost all the results show a superior performance than the chance level.

inner loop were retrained on the entire outer training set and evaluated on the held-out subjects. More details about these classifiers, grid search, and hyperparameters can be found in the Supplementary Materials.

V. DISCUSSION

Our analysis revealed that (i) increasing the number of physiological modalities improves user performance classification, and (ii) reliable estimates can be achieved with a carefully selected subset of modalities, highlighting the importance of signal quality and task relevance over sheer signal count.

Across 3 visualization types and 120 participants, the Easy condition yielded higher task accuracy rates and faster response times, while relevant markers, such as the ones derived from ET devices, tracked these differences reliably. EEG

primarily reflected workload through increases in Spectral entropy, while EDA showed strongest effects for peak-based dynamics.

All classifiers were able to discriminate Hard from Easy trials significantly above chance level and for every physiological modality (see Figure 8). When combining all modalities, results were on par with those obtained with the behavioural metrics (Figure 7). We note that, in real-time applications, continuous implicit monitoring of behavioural data is only available until task completion, hence making the combination of physiological signals a practical alternative. Together, our findings support the use of multimodal sensing for estimating user performance in InfoVis and motivate future adaptive interfaces that can select among alternative visual encodings based on user state. PERCEIV does not evaluate such a closed-loop system; instead, it provides empirical evidence about responses to controlled encoding variants, complementing recent system-level work where neurophysiological workload estimates are used to dynamically adapt visual dashboards in real time [56].

A. Use Cases and Applications

Our analysis suggests that user performance can be reliably estimated from simple behavioural signals (e.g., RT and accuracy) and further supported by multimodal physiology (e.g., EEG, EDA, and ET). In practice, this enables two complementary integration paths. One is behaviour-gated adaptation, where fast, high-signal behavioural features act as primary triggers, while physiology provides confirmation and context. For example, when RTs drift upward and accuracy declines, concurrent increases in EEG theta power and entropy, or rises in pupil dilation and fixation count, may reveal that the user is operating under high load. The interface can then simplify charts, surface task-relevant encodings, or defer non-critical notifications until load subsides. Because behaviour is already being produced, such adjustments can happen with very low latency and minimal false positives.

The second path, equally important and emphasised in the research literature, is physiology-first adaptation [4], [79], where no overt behaviour is required or even available; cf. the so-called *cold start problem* [80]. Physiological measurements provide a continuous window into latent cognitive dynamics and can be acted upon before a user makes an error or even engages with a control. In PERCEIV, tonic increases in theta power and entropy, together with elevated fixation values or EDA-AUC, frequently preceded performance costs in the Hard condition. These dynamics make it possible to proactively adjust visualisation density, cue useful summaries, or offer alternative question scaffolds while the user is still reading or orienting. In supervisory or safety-critical contexts, the ability to issue early soft alerts based solely on physiology is a clear advantage over purely behavioural schemes, which must wait for clicks or answers to accumulate.

In real-world deployments, we would recommend a hybrid loop: use physiology to maintain a running estimate of moment-to-moment workload, fall back to behaviour when available to recalibrate and disambiguate, and fuse both to decide when, how, and by how much to adapt. To support

transparency and auditability, adaptations should be conservative, and be logged alongside summary indices rather than raw bio-signals. When behaviour is absent (e.g., during prolonged viewing, reading, or briefing), physiology alone can still drive low-risk helpful behaviours such as highlighting, progressive disclosure, and pacing suggestions, and hand off to behaviour-gated control once interaction resumes.

Together, these capabilities create new opportunities to advance research in several areas, including: (i) intelligent systems across domains such as entertainment, e-health, e-learning, and e-government, (ii) enhancing the effectiveness of Human-AI collaboration, (iii) developing risk assessment and management systems for robotics and high-risk occupations, and (iv) supporting personal applications aimed at improving health-related quality of life and overall well-being.

B. Limitations and Recommendations

In the PERCEIV dataset, we prioritised internal validity and breadth across visualisation types over ecological breadth. Workload labels combine task complexity with stimulus load, which captures the regimes of interest but limits causal separation of those factors. Future iterations should decouple these variables, or introduce graded levels within each factor (instead of binary easy/hard splits). Our feature set was intentionally kept pragmatic, i.e: frontal EEG subsets or compact connectivity metric (theta PLV), leaving room for richer spatial coverage, band-specific connectivity information, and gaze features tied to visualization AOIs or scanpaths.

In addition, visualization type was assigned between participants rather than within participants. Therefore, the dataset supports analyses of Easy–Hard effects within each visualization group, but apparent differences in effect-size magnitude across visualization conditions should be interpreted descriptively unless modeled with appropriate between-subjects statistics.

Another limitation concerns the domain specificity of the stimuli. All tasks were framed around criminal-network scenarios, which provided a controlled and coherent context across visualization types but may also influence semantic interpretation and perceived task relevance. Therefore, generalization to other applied domains, such as medical, financial, or operational dashboards, should be made cautiously and validated in future datasets with broader thematic coverage.

A further limitation is the cross-site EEG heterogeneity. Although the experimental protocol was shared across sites and the main Easy–Hard EEG effects were broadly consistent after within-site normalization, Site 1 used a 32-channel gel-based system whereas Site 2 used a 12-channel dry-electrode system. Therefore, we recommend that future users consider site-specific analyses or include site/device as a covariate when modeling EEG features [72].

In addition, it is important to acknowledge that while hand-crafted features are grounded in strong prior hypotheses [3], [11], they may also discard potentially informative signal components. Notably, a growing number of recent studies are shifting toward end-to-end Deep Learning (DL) approaches [81], [82], [83], in which discriminative features are automatically

discovered in a data-driven manner, an exploitation path that remains to be explored with the PERCEIV dataset.

We also note ethical and operational considerations that accompany physiology-driven studies, due to the direct biological connection of the user to the computer. Wherever possible, (i) processing should occur on-device with only derived indices retained for analysis, (ii) users should be able to opt out, and (iii) models should be monitored for stability and fairness across demographics. From a design standpoint, behaviour and physiology should be fused in a way that privileges interpretability and user control. Behaviour remains the strongest, lowest-cost indicator when available, but neurophysiology offers anticipatory, bias-resistant signals when behaviour is sparse, absent or cannot account for effects not consciously available to the user. Finally, we encourage the community to treat this dataset as a foundation to study sequential decision-making processes; e.g., policies that weigh adaptation benefits against cognitive costs.

C. Future Work

First and foremost, we plan to broaden task coverage and labels, adding independent manipulations of complexity and time pressure, also considering longitudinal sessions. Because we release all raw data and stimuli, it is possible to introduce multi-level difficulty factors. For example, additional question tiers could require integrating three or more attributes, enabling finer-grained cognitive demand manipulations beyond the Easy/Hard splits we used. In addition, we plan to expand the signal space with richer eye-movement descriptors (e.g., saccade kinematics, scanpath similarity), and spatially resolved EEG features, together with improved artifact handling and synchronized quality flags to facilitate robust real-time use.

VI. CONCLUSION

We have presented a large-scale multimodal dataset and analysis for studying cognitive load features in adaptive InfoVis interfaces. Across three visualization types (network graphs, distribution charts, and timelines) and multiple strategies that change visual encodings, we showed converging evidence that adaptive visualizations improve both accuracy and response speed. By releasing synchronized data at the trial level, together with software code and reproducible analyses, we aim to provide a common test-bed for cognitive-state’s features estimation in adaptive visualization tasks. We also aim to accelerate progress towards the development of more transparent, trustworthy neuro-adaptive systems that improve decision quality and user experience.

REFERENCES

- [1] L. Perkhof, O. M. Lehner, and H. Treiblmaier, “Infovis: The impact of information overload on decision making outcome in high complexity settings,” in *Proc. AIS SIGHCI Pre-ICIS Workshop*, 2016.
- [2] Y. Wang, W. Wang, A. Abdelhafez, M. Elfars, Z. Hu, M. Bâce, and A. Bulling, “Salchartqa: Question-driven saliency on information visualisations,” in *Proc. CHI*, 2024.
- [3] S. H. Fairclough, *Current research in neuroadaptive technology*. Academic Press, 2022, ch. Designing human-computer interaction with neuroadaptive technology.

- [4] L. M. Hirshfield, R. Gulotta, S. Hirshfield, S. Hincks, M. Russell, R. Ward, T. Williams, and R. Jacob, "This is your brain on interfaces: enhancing usability testing with functional near-infrared spectroscopy," 2011.
- [5] E. M. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang, "Using fnirs brain sensing to evaluate information visualization interfaces," in *Proc. CHI*, 2013.
- [6] S. Haddad, K. Latifzadeh, S. Duraisamy, J. Vanderdonckt, O. Daassi, S. Belghith, and L. A. Leiva, "Good guis, bad guis: Affective evaluation of graphical user interfaces," in *Proc. UMAP*, 2024.
- [7] T. K. Fredericks, S. D. Choi, J. Hart, S. E. Butt, and A. Mital, "An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads," *Int. J. Ind. Ergon.*, vol. 35, no. 12, 2005.
- [8] M. Shakouri, L. H. Ikuma, F. Aghazadeh, and I. Nahmens, "Analysis of the sensitivity of heart rate variability and subjective workload measures in a driving simulator: the case of highway work zones," *Int. J. Ind. Ergon.*, vol. 66, 2018.
- [9] M. Haak, S. Bos, S. Panic, and L. J. Rothkrantz, "Detecting stress using eye blinks and brain activity from EEG signals," *Proc. DCII*, vol. 35, 2009.
- [10] M. ElKomy, Y. Abdelrahman, M. Funk, T. Dingler, A. Schmidt, and S. Abdennadher, "Abbas: an adaptive bio-sensors based assistive system," in *Proc. CHI EA*, 2017.
- [11] L. A. Leiva, I. Arapakis, M. Vakalellis, and J. Vanderdonckt, "Context-aware adaptive visualizations for critical decision making," 2023.
- [12] K. Gramann, S. H. Fairclough, T. O. Zander, and H. Ayaz, "Trends in neuroergonomics," *Front. Hum. Neurosci.*, vol. 11, 2017.
- [13] E. Ratti, S. Waninger, C. Berka, G. Ruffini, and A. Verma, "Comparison of medical and consumer wireless EEG systems for use in clinical trials," *Front. Hum. Neurosci.*, vol. 11, 2017.
- [14] C. Cannard, H. Wabbeh, and A. Delorme, "Validating the wearable muse headset for EEG spectral analysis and frontal alpha asymmetry," in *Proc. BIBM*, 2021.
- [15] J. K. Nuamah, Y. Seong, S. Jiang, E. Park, and D. Mountjoy, "Evaluating effectiveness of information visualizations using cognitive fit theory: A neuroergonomics approach," *Appl. Ergon.*, vol. 88, 2020.
- [16] E. W. Anderson, K. C. Potter, L. E. Matzen, J. F. Shepherd, G. A. Preston, and C. T. Silva, "A user study of visualization effectiveness using EEG and cognitive load," *Comput. Graph. Forum*, vol. 30, no. 3, 2011.
- [17] M. Khateeb, S. M. Anwar, and M. Alnowami, "Multi-domain feature fusion for emotion classification using deap dataset," *IEEE Access*, vol. 9, 2021.
- [18] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through EEG and ecg signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 1, 2018.
- [19] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, 2018.
- [20] M. Mustafa, L. Lindemann, and M. Magnor, "EEG analysis of implicit human visual perception," in *Proc. CHI*, 2012.
- [21] H. B. S. Consortium, E. Aarts, A. Akkerman, M. Altgassen, R. Bartels, D. Beckers, K. Bevelander, E. Bijleveld, E. Blaney Davidson, A. Boleij *et al.*, "Protocol of the healthy brain study: An accessible resource for understanding the human brain and how it dynamically and individually operates in its bio-social context," *PLOS ONE*, vol. 16, no. 12, 2021.
- [22] I. Vessey, "Cognitive fit: A theory-based analysis of the graphs versus tables literature," *Decis. Sci.*, vol. 22, no. 2, 1991.
- [23] K. Mohanavelu, S. Poonguzhali, D. Ravi, P. K. Singh, M. Mahajabin, K. Ramachandran, U. K. Singh, and S. Jayaraman, "Cognitive workload analysis of fighter aircraft pilots in flight simulator environment," *Def. Sci. J.*, vol. 70, no. 2, 2020.
- [24] E. Johannessen, A. Szulewski, N. Radulovic, M. White, H. Braund, D. Howes, D. Rodenburg, and C. Davies, "Psychophysiological measures of cognitive load in physician team leaders during trauma resuscitation," *Comput. Hum. Behav.*, vol. 111, 2020.
- [25] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger, "Using psycho-physiological measures to assess task difficulty in software development," in *Proc. ICSE*, 2014.
- [26] M. Chang, Y. Wang, H. W. Wang, Y. Zhou, A. Bulling, and C. X. Bearfield, "Tell me without telling me: Two-way prediction of visualization literacy and visual attention," 2025.
- [27] G. Cartella, M. Cornia, V. Cuculo, A. D'Amelio, D. Zanca, G. Boccignone, G. Boccignone, and R. Cucchiara, "Trends, applications, and challenges in human attention modelling," in *Proc. IJCAI*, 2024.
- [28] D. Shi, Y. Wang, Y. Bai, A. Bulling, and A. Oulasvirta, "Chartist: Task-driven eye movement control for chart reading," 2025.
- [29] T. Blascheck and T. Ertl, "Towards analyzing eye tracking data for evaluating interactive visualization systems," in *Proc. BELIV*, 2014.
- [30] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister, "What makes a visualization memorable?" *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, 2013.
- [31] J. Heer and M. Bostock, "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," in *Proc. CHI*, 2010.
- [32] A. Lopez-Cardona, P. Emami, S. Idesis, S. Duraisamy, L. A. Leiva, and I. Arapakis, "A comparative study of scanpath models in graph-based visualization," in *Proc. ETRA*, 2025.
- [33] W. S. Cleveland and R. McGill, "Graphical perception: Theory experimentation, and application to the development of graphical methods," *J. Amer. Stat. Assoc.*, vol. 79, no. 387, 1984.
- [34] S. M. Kosslyn, "Understanding charts and graphs," *Appl. Cognit. Psychol.*, vol. 3, no. 3, 1989.
- [35] S. Pinker, *A theory of graph comprehension*. Psychology Press, NY, 1990.
- [36] E. R. Tufte, *The Visual Display of Quantitative Information*. Graphics Press, 2001.
- [37] J. Hullman, E. Adar, and P. Shah, "Benefitting infovis with visual difficulties," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, 2011.
- [38] Z. Zeng and L. Battle, "A review and collation of graphical perception knowledge for visualization recommendation," in *Proc. CHI*, 2023.
- [39] M. Salous, D. Küster, K. Scheck, A. Dikfidan, T. Neumann, F. Putze, and T. Schultz, "Smarthelm: User studies from lab to field for attention modeling," in *Proc. SMC*, 2022.
- [40] P. Toreini, M. Langner, A. Maedche, S. Morana, and T. Vogel, "Designing attentive information dashboards," *J. Assoc. Inf. Syst.*, vol. 23, no. 2, 2022.
- [41] A. J. Karan, T. Demazure, P.-M. Leger, E. Labonte-LeMoyne, S. Senecal, M. Fredette, and G. Babin, "Toward a hybrid passive BCI for the modulation of sustained attention using EEG and fnirs," *Front. Hum. Neurosci.*, vol. 13, 2019.
- [42] H. Aghajani, M. Garbey, and A. Omurtag, "Measuring mental workload with eeg+ fnirs," *Front. Hum. Neurosci.*, vol. 11, 2017.
- [43] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psychophysiological measures for assessing cognitive load," in *Proc. UbiComp*, 2010.
- [44] M. Züger and T. Fritz, "Interruptibility of software developers and its prediction using psycho-physiological sensors," in *Proc. CHI*, 2015.
- [45] T. A. D'Amelio, L. A. Galán, E. A. Maldonado, A. A. D. Barquinero, J. R. Cuello, N. M. Bruno, E. Tagliazucchi, and D.-A. Engemann, "Emotion recognition systems with electrodermal activity: From affective science to affective computing," *Neurocomputing*, vol. 651, 2025.
- [46] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond memorability: Visualization recognition and recall," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, 2016.
- [47] D. Toker, C. Conati, B. Steichen, and G. Carenini, "Individual user characteristics and information visualization: connecting the dots through eye tracking," in *Proc. CHI*, 2013.
- [48] P. Polatsek, M. Waldner, I. Viola, P. Kapec, and W. Benesova, "Exploring visual attention and saliency modeling for task-based visual analysis," *Comput. Graph.*, vol. 72, 2018.
- [49] N. W. Kim, Z. Bylinskii, M. A. Borkin, K. Z. Gajos, A. Oliva, F. Durand, and H. Pfister, "Bubbleview: An interface for crowdsourcing image importance maps and tracking visual attention," *ACM Trans. Comput.-Hum. Interact.*, vol. 24, no. 5, 2017.
- [50] M. Gjoreski, T. Kolenik, T. Knez, M. Luštrek, M. Gams, H. Gjoreski, and V. Pejović, "Datasets for cognitive load inference using wearable sensors and psychological traits," *Appl. Sci.*, vol. 10, no. 11, 2020.
- [51] A. Bhatti, P. Angkan, B. Behinaein, Z. Mahmud, D. Rodenburg, H. Braund, P. J. Mclellan, A. Ruberto, G. Harrison, D. Wilson *et al.*, "Clare: Cognitive load assessment in realtime with multimodal data," *IEEE Trans. Cogn. Dev. Syst.*, 2025.
- [52] A. Kalatzis, A. Teotia, V. G. Prabhu, and L. Stanley, "A database for cognitive workload classification using electrocardiogram and respiration signal," in *Proc. AHFE*, 2021.
- [53] M. P. Oppelt, A. Foltyn, J. Deuschel, N. R. Lang, N. Holzer, B. M. Eskofier, and S. H. Yang, "Adabase: a multimodal dataset for cognitive load estimation," *Sensors*, vol. 23, no. 1, 2022.
- [54] V. Markova, T. Ganchev, and K. Kalinkov, "Clas: A database for cognitive load, affect and stress recognition," in *Proc. BIA*, 2019.
- [55] V. Yoghoudjian, Y. Yang, T. Dwyer, L. Lawrence, M. Wybrow, and K. Marriott, "Scalability of network visualisation from a cognitive load perspective," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, 2020.

- [56] A. Lopez-Cardona, M. M. Bruns, N. T. Attygalle, S. Idesis, M. Salvatori, K. Raftopoulos, K. Oikonomou, S. Duraisamy, P. Emami, N. Latreche *et al.*, “Context-aware adaptive visualizations for critical decision making,” *Proc. ECAI*, 2025.
- [57] J. W. Peirce, “Psychopy—psychophysics software in python,” *J. Neurosci. Methods*, vol. 162, no. 1-2, 2007.
- [58] K. Latifzadeh and L. A. Leiva, “Gustav: Cross-device cross-computer synchronization of sensory signals,” in *Adj. Proc. UIST*, 2022.
- [59] F. Marini, C. Lee, J. Wagner, S. Makeig, and M. Gola, “A comparative evaluation of signal quality between a research-grade and a wireless dry-electrode mobile EEG system,” *J. Neural Eng.*, vol. 16, no. 5, 2019.
- [60] I. Chanpornpakdi, R. Mizuochi, and M. G. Machizawa, “Evaluation of influence of positions and numbers of EEG electrodes on quantification of independent component matrix,” in *Proc. APSIPA ASC*, 2022.
- [61] K. Rayner, “Eye movements and attention in reading, scene perception, and visual search,” *Q. J. Exp. Psychol.*, vol. 62, no. 8, 2009.
- [62] G. Wu, L. Bolliger, D. Reich, and L. Jäger, “An eye opener regarding task-based text gradient saliency,” in *Proc. CMCL*, 2024.
- [63] H. C. Cuve, J. Stojanov, X. Roberts-Gaal, C. Catmur, and G. Bird, “Validation of gaze-point low-cost eye-tracking and psychophysiology bundle,” *Behav. Res. Methods*, vol. 54, no. 2, 2022.
- [64] G. Koutroumpas, M. Mazzini, S. Idesis, M. Masias Bruns, J. M. Jose, S. Abadal, and I. Arapakis, “Beyond clicks: Eye-tracking insights into user responses to different recommendation types,” in *Proc. ICTIR*, 2025.
- [65] A. Lopez-Cardona, S. Idesis, and I. Arapakis, “Integrating cognitive processing signals into language models: A review of advances, applications and future directions,” 2025.
- [66] A. Lopez-Cardona, S. Idesis, M. Barreda-Ángeles, S. Abadal, and I. Arapakis, “Oasst-etc dataset: Alignment signals from eye-tracking analysis of llm responses,” *Proc. ACM Hum.-Comput. Interact.*, vol. 9, no. 3, 2025.
- [67] W. T. Cochran, J. W. Cooley, D. L. Favin, H. D. Helms, R. A. Kaenel, W. W. Lang, G. C. Maling, D. E. Nelson, C. M. Rader, and P. D. Welch, “What is the fast fourier transform?” *Proc. IEEE*, vol. 55, no. 10, 1967.
- [68] W.-L. Zheng and B.-L. Lu, “Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks,” *IEEE Trans. Auton. Ment. Dev.*, vol. 7, no. 3, 2015.
- [69] S. Chikhi, N. Matton, and S. Blanchet, “EEG power spectral measures of cognitive workload: A meta-analysis,” *Psychophysiology*, vol. 59, no. 6, 2022.
- [70] G. Koutroumpas, M. Masias Bruns, S. Idesis, M. Mazzini, J. M. Jose, S. Abadal, and I. Arapakis, “Organic vs. synthetic attention: Evaluating the utility of eye tracking augmentation in recommender systems,” *Proc. ACM Comput. Graph. Interact. Tech.*, vol. 9, no. 2, 2026.
- [71] S. Idesis, M. Masias Bruns, P. Emami, S. Duraisamy, L. A. Leiva, and I. Arapakis, “Evaluating the sensitivity of dry and gel-based wearable eeg for cognitive load estimation,” *Biomedical Physics & Engineering Express*, vol. 12, no. 3, p. 035031, 2026.
- [72] —, “Evaluating the sensitivity of dry and gel-based wearable EEG for cognitive load estimation,” *bioRxiv*, 2026.
- [73] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
- [74] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. John Wiley & Sons, 2013.
- [75] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, 1995.
- [76] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, 2001.
- [77] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Ann. Stat.*, vol. 29, no. 5, 2001.
- [78] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, 1967.
- [79] M. A. Webster and K. E. Tregillus, “Visualizing visual adaptation,” *J. Vis. Exp.*, no. 122, 2017.
- [80] B. Lika, K. Kolomvatsos, and S. Hadjiefthymiades, “Facing the cold start problem in recommender systems,” *Expert Syst. Appl.*, vol. 41, no. 4, 2014.
- [81] M. A. Khan, H. Asadi, M. R. Chalak Qazani, A. Arogbonlo, S. Nahavandi, and C. P. Lim, “Predicting cognitive load in immersive driving scenarios with a hybrid CNN-RNN model,” 2024.
- [82] M. Safari, R. Shalbaf, S. Bagherzadeh, and A. Shalbaf, “Classification of mental workload using brain connectivity and machine learning on electroencephalogram data,” *Sci. Rep.*, vol. 14, no. 1, 2024.
- [83] G. Siddhad, P. P. Roy, and B.-G. Kim, “Neural networks meet neural activity: Utilizing EEG for mental workload estimation,” in *Proc. ICPR*, 2024.