Decoding Cognitive Load: Eye-Tracking Insights into Working Memory and Visual Attention

Xiaofu Jin
Academy of Interdisciplinary Study
The Hong Kong University of Science and Technology
Hong Kong SAR, China
xjinao@connect.ust.hk

Yunpeng Bai
National University of Singapore
Singapore, Singapore
bai_yunpeng99@u.nus.edu

Lina Xu
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
lxu582@connect.hkust-gz.edu.cn

Shuai Ma
Aalto University
Helsinki, Finland
shuai.ma@aalto.fi

Danqing Shi
Aalto University
Helsinki, Finland
danqing.shi@aalto.fi

Luwen Yu
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
luwenyu@hkust-gz.edu.cn

Mingming Fan
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
mingmingfan@ust.hk

ABSTRACT

Understanding and classifying cognitive workload is a critical challenge in educational technology and human-computer interaction (HCI). While cognitive load is often treated as a single, generalized concept, its nuanced components—such as working memory load and visual attention load—play distinct roles in learning environments. To investigate these differences, we conducted a controlled experiment, collecting a comprehensive eye-tracking dataset comprising 528,017 data points across varied cognitive tasks. Leveraging machine learning, we demonstrate that these cognitive states can be classified, revealing measurable distinctions between load types. Our findings pave the way for adaptive learning systems that dynamically tailor instructional content based on cognitive state assessments. This research contributes to the development of personalized, AI-enhanced educational tools, advancing both theoretical understanding and practical applications of eye-tracking in education, cognitive assessment, and HCI.

CCS CONCEPTS

• **Computing methodologies → Machine learning approaches**; • **Human-centered computing → HCI theory, concepts and models**; **User models**; **Laboratory experiments**.

ETRA '25, May 26–29, 2025, Tokyo, Japan

ACM ISBN 979-8-4007-1487-0/2025/05
https://doi.org/10.1145/3715669.3725864

KEYWORDS

eye tracking, cognitive load, working memory, visual attention

ACM Reference Format:
Xiaofu Jin, Yunpeng Bai, Lina Xu, Shuai Ma, Danqing Shi, Luwen Yu, and Mingming Fan. 2025. Decoding Cognitive Load: Eye-Tracking Insights into Working Memory and Visual Attention. In *2025 Symposium on Eye Tracking Research and Applications (ETRA '25), May 26–29, 2025, Tokyo, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3715669.3725864

1 INTRODUCTION

Measuring cognitive workload is essential for designing effective and adaptive interactive systems, particularly in educational contexts where understanding learners' cognitive states can enhance teaching and learning outcomes. Traditional methods like NASA-TLX [Hart 2006; Hart and Staveland 1988] rely on subjective self-reporting and often fail to capture real-time cognitive states [Guastello et al. 2015]. While advanced technologies such as fMRI [Engel et al. 1994], EEG [Gevins and Smith 2003], and fNIRS [Yuksel et al. 2016] provide more precise measurements, they are expensive and impractical for widespread use in educational settings. Eye-tracking technology offers a promising alternative due to its non-invasive and cost-effective nature, making it suitable for assessing cognitive workload in real-time across both in-person and remote learning environments.

Eye-tracking has broad applications across fields such as driving [Gomaa et al. 2022], healthcare [Naik et al. 2022], and digital design [Wang et al. 2014]. In education, it is becoming a valuable tool for understanding learning processes and adapting instructional methods. However, most existing research focuses on overall task difficulty (e.g., low, medium, high) and treats cognitive load as a single construct [Duchowski et al. 2020; Marshall 2002; Orden et al. 2001; Rudmann et al. 2003]. This generalization overlooks the distinction between different types of cognitive load, such as working memory and visual attention [Kosch et al. 2023]. Differentiating these components is crucial because working memory involves maintaining and manipulating information, while visual attention refers to focusing on relevant stimuli. Understanding these differences allows for more precise adaptation in educational settings, such as reducing memory strain or improving attention management in complex tasks.

To address these gaps, we conducted a controlled experiment to distinguish between working memory load and visual attention load in educational contexts. Using eye-tracking technology, we collected data from tasks designed to induce varying levels of these cognitive demands and applied machine learning models to classify these states. Our findings demonstrate that eye-tracking data can reveal fine-grained distinctions in cognitive load, offering practical insights for designing adaptive learning environments.

Differentiating cognitive load types is also critical in other high-stakes domains. In healthcare, for instance, cognitive overload can impair performance during complex procedures like surgery [Cicourel 2004; Zenati et al. 2019]. Our approach could enable systems to dynamically adjust workflows by highlighting essential information during visual overload or simplifying processes during memory strain. Similarly, in immersive learning environments such as virtual reality (VR), adaptive systems could respond to cognitive load in real-time, improving training effectiveness in areas like emergency response [Marshall 2002; Stoeve et al. 2022].

In summary, we make the following contributions:

- **Dataset**: We developed the first eye-tracking dataset (528,017 data points) that captures distinct working memory and visual attention loads under controlled task conditions. This dataset addresses a gap in available resources, supporting further research in cognitive load analysis and its applications in education.
- **Methodology**: We pioneered a novel approach to distinguishing meta-level cognitive loads by leveraging diverse eye-tracking metrics and machine learning models. This work lays a foundation for future research on cognitive load measurement, particularly in educational settings.
- **Applications**: By exploring the feasibility of fine-grained cognitive load classification, our study offers actionable insights for designing adaptive technologies in domains such as personalized learning, educational gaming, and virtual reality-based training.

2 RELATED WORK

2.1 Eye Tracking Measure

In this subsection, we summarize the current major metrics that researchers have used to indicate cognitive workload through eye-tracking activities.

2.1.1 Pupil Measure. Pupil-based metrics offer a non-invasive, continuous, and unobtrusive means of assessing cognitive workload. Previous work has shown that pupil diameter correlates between pupil dilation and problem difficulty in a range of activities [Beatty 1982; Chen et al. 2011; Kahneman 1973], including visual search tasks [Marshall 2002; Rudmann et al. 2003], serious game [Calà et al. 2024] and driving activities [Bozkir et al. 2019]. However, traditional pupil diameter measurements are limited by sensitivity to off-axis distortion [Krejtz et al. 2018] and ambient light changes [Duchowski et al. 2018], which can affect the reliability of cognitive load assessments. To address these limitations, Duchowski et al. introduced two frequency-based indices: the Index of Pupillary Activity (IPA) [Duchowski et al. 2018] and the Low/High Index of Pupillary Activity (LHIPA) [Duchowski et al. 2020]. These indices focus on dynamic changes in pupil diameter rather than absolute measurements, making them less susceptible to distortions caused by viewing angle and ambient light conditions.

2.1.2 Fixations. Fixations (voluntary eye movements focusing on a specific area) reflect cognitive load through attentional allocation and information processing mechanisms. Longer fixation durations typically indicate deeper cognitive engagement, such as when integrating complex information or resolving task conflicts, as seen in visuospatial memory tasks [Orden et al. 2001]. This occurs because increased cognitive strain prolongs the time needed to encode or interpret stimuli [Chen et al. 2011; Saeedpour-Parizi et al. 2020]. Conversely, a higher fixation rate (more frequent fixations on an area of interest) often signals active searching or difficulty locating task-relevant information, suggesting elevated attentional demands [Rudmann et al. 2003]. For example, in problem-solving tasks, learners may exhibit clustered fixations on critical elements when struggling to reconcile concepts, directly linking gaze behavior to working memory load [Fadardi et al. 2022; Holmqvist et al. 2011].

2.1.3 Eye Blinks. Eye blinks serve as a natural indicator of perceived workload, with involuntary blinks signaling cognitive fatigue. Research has consistently linked lower blink rates and longer blink latencies to increased cognitive load [Chen and Epps 2014; Chen et al. 2011; Orden et al. 2001]. This relationship suggests that as cognitive demands rise, users blink less frequently and exhibit longer intervals between blinks, making blink metrics useful for gauging mental effort.

2.1.4 Saccades. Saccades refer to the rapid eye movements that shift focus between fixations (areas of interest), typically occurring within 25 to 60 milliseconds [Perego et al. 2012]. Saccadic metrics, such as speed, have been found to be reliable indicators of cognitive workload. Faster or larger saccades are commonly observed in high-load conditions [Chen et al. 2011], particularly in contexts like driving, where increased cognitive demands are reflected in greater saccadic velocities [Biswas and Prabhakar 2018].

These eye-tracking measures provide a comprehensive, real-time assessment of cognitive workload during learning tasks. By integrating pupil dynamics, fixations, blinks, and saccades, researchers can objectively quantify mental effort without interrupting task performance. This is particularly valuable in educational settings, where understanding cognitive load can help optimize instructional design, adapt task difficulty, and identify struggling learners

2.2 Differentiating Cognitive Load States

Cognitive Load Theory (CLT) identifies three types of cognitive load: intrinsic, extraneous, and germane [Sweller et al. 1998]. These types describe mental demands during task performance and their relationship to cognitive functions like working memory and visual attention. Intrinsic load arises from task complexity, taxing working memory, which holds and manipulates information [Baddeley 1992]. High intrinsic load can overwhelm working memory, leading to errors and reduced efficiency. Extraneous load stems from inefficient information presentation, often straining visual attention, especially in cluttered environments, leading to cognitive fatigue and lower performance [Ranchet et al. 2019]. Germane load focuses on learning and schema development, engaging both working memory and visual attention [Sweller 2010].

Recognizing the distinct roles of working memory and visual attention in cognitive load management informs better system design. The n-back task, a widely used method in cognitive psychology, requires participants to recall whether a current stimulus (e.g., letters, numbers, or images) matches one presented n steps earlier in a sequence (e.g., 1-back or 2-back). This task systematically varies working memory load, making it a valuable tool for studying intrinsic cognitive load in learning contexts. For example, higher n-back levels (e.g., 3-back) mimic the demands of retaining intermediate steps in math problems or synthesizing lecture content, while lower levels (1-back) resemble rote memorization. Insights from n-back studies, such as the impact of load on performance and mental effort [Bacchin et al. 2023; Duchowski et al. 2020, 2018], can inform educational strategies like chunking complex information or adaptive task sequencing to optimize cognitive load in classrooms [Sweller 2010]. For tasks involving visual attention, the visual search task is a prominent method used in HCI to study extraneous cognitive load. In such tasks, users must locate targets within a visual field, often reflecting real-world scenarios. For example, participants might eliminate randomly appearing colored cubes by pointing and clicking with a laser-pointer controller [Szczepaniak et al. 2024]. Similarly, browsing online shopping websites to find specific items demonstrates how interface design influences visual search performance [Wang et al. 2014]. These studies highlight the importance of reducing extraneous load through strategies such as decluttering interfaces and improving the salience of key elements [Wickens et al. 2021].

3 METHOD

To examine cognitive load through eye-tracking, we collected data across tasks with varying working memory and visual attention demands. This section details the collection of our dataset, providing an overview of the participants, data collection methods, and final dataset composition after quality control.

3.1 Hypotheses

Prior to designing the experiment, we formulated the following hypotheses based on cognitive load theory and prior eye-tracking research: (1) Working memory task: higher n-back levels (1-back, 2-back) will induce greater cognitive load, reflected in increased pupil diameter, longer fixation durations, and reduced blink rates compared to the 0-back baseline. (2) Visual search task: larger distractor sets (15 vs. 5) will increase extraneous cognitive load, manifesting as higher saccadic velocities and more fixations due to heightened visual attention demands. (3) Working memory load (intrinsic) and visual search difficulty (extraneous) will exhibit distinct eye tracking signatures, allowing the assessment of granular workload. These hypotheses guided our task design and metric selection (e.g., pupil metrics for intrinsic load, saccades for extraneous load).

3.2 Participant Recruitment

Participants were recruited through social media channels targeting students and staff members. From 48 initial respondents, we selected 22 participants who met our eligibility criteria: normal or corrected-to-normal vision, no history of cognitive impairments, and availability during scheduled testing sessions. During data analysis, one participant was excluded for performing below our predetermined accuracy threshold (40% correct responses), suggesting either difficulty understanding the task or non-compliance with instructions.

3.3 Experimental Setting and Apparatus

A Tobii Pro Nano Eye-Tracker was used to record eye movements binocularly at a sampling rate of 60 Hz. The participants' positions were fixed throughout the experimental procedure to ensure consistent eye-tracking with 57 cm. Each participant underwent a calibration process with a visual angle of 0.5 degrees. Visual stimuli were displayed on a computer monitor with a resolution of 1920×1080 pixels. The entire procedure was automated using HTML, with minimal external intervention. Participant responses were collected using a standard mouse connected to the presentation computer, which was positioned on the participant's dominant hand side. The experiment took place in a laboratory without windows to minimize ambient light, maintaining controlled lighting conditions throughout the study.

3.4 Experimental Procedure

We implemented three levels of working memory and visual search tasks, drawing on traditional psychological paradigms [Stoet 2010]. For the working memory task, we adopted the standard n-back task, commonly used in cognitive neuroscience [Kane et al. 2007]. In this task, participants are required to decide whether each stimulus in a sequence matches the one presented n items ago. Based on the results of two pilot studies with five participants each, we found that the 3-back task was too difficult for participants, leading to low accuracy and signs of cognitive overload. Therefore, we adjusted the task to include 0-back, 1-back, and 2-back levels for the working memory task. In the 0-back condition, participants were instructed to determine whether the current letter was "H." In the 1-back condition, they determined if the letter matched the one immediately before, and in the 2-back condition, they compared it to the letter presented two items ago. Eight phonologically distinct letters (B, F, K, H, M, Q, R, X) served as stimuli, with each letter appearing once per block. An additional four letters acted as targets, creating blocks with four targets (33% of trials) and eight foils (67% of trials). To prevent participants from relying on perceptual features, the letters were presented randomly in either uppercase or lowercase. Each trial began with a stimulus presentation lasting 500 ms, followed by a blank screen for a 2,000 ms interstimulus interval. Participants pressed the left mouse button for "yes" and the right mouse button for "no." Visual stimuli were used instead of auditory stimuli to better simulate real-world tasks performed on screens. For the visual search task, we used a task provided by PsyToolkit [Stoet 2010], which simulates the process of searching for a target icon or feature on a screen—an important and frequent activity in everyday life. Participants were tasked with finding a target icon  among varying numbers of distractors (set to 5, 10, or 15 distractors based on pilot study findings). Each visual search trial began with the stimulus presentation for 2,000 ms, followed by a 500 ms blank screen to align the trial duration with that of the working memory task. Each task (N-back and visual search) was repeated, with repetitions conducted consecutively (e.g., N-back followed by N-back, and visual search followed by visual search). The order of task blocks (N-back vs. visual search) was counterbalanced using a balanced Latin square design to minimize potential order effects. Within each task type, the difficulty levels were randomized to ensure no systematic bias within tasks. The experimental design incorporated three categories of variables: the independent variables consisted of task type (working memory n-back task vs. visual search task) and difficulty levels (0/1/2-back conditions for the working memory task, and 5/10/15 distractor conditions for the visual search task); the dependent variables included eye-tracking metrics (pupil diameter, fixation duration, saccadic velocity) and behavioral performance measures (task accuracy, reaction time); while control variables accounted for individual differences (assessed through Digit Span and Stroop tests) and environmental factors (standardized lighting conditions, fixed viewing distance), ensuring the reliability of experimental results.

Participants were briefed on the experimental procedure, informed of the task details and risks, and asked to sign a consent form if they agreed to participate. Demographic information, including age, sex, education level, working memory capacity (measured using the Digit Span Task)[Woods et al. 2011], and visual attention capacity (measured using the Stroop test)[Stroop 1935], was collected. Before the formal experiment, participants completed a familiarization phase, which included 24 practice trials for each task.

3.5 Dataset Information

After excluding data from one participant due to quality issues, our final dataset includes 21 participants, with 13 males and 8 females. Each participant completed six conditions, with eye-tracking data recorded at 60 Hz for one minute per condition. To ensure consistency, we down-sampled the relaxation condition to match the size of other conditions. After data cleaning, including the removal of missing data, each participant has approximately 3,600 data points. The entire dataset for all seven conditions includes 528,017 data points, with 75,431 per condition. In this study, we group the three levels of working memory and visual attention into separate categories to simplify the problem and focus on distinguishing between the two types. The dataset will be open-sourced in the future.

4 RESULTS

We describe the formation of eye-tracking features. We start with index features derived from existing literature as we described in Sec. 2.1. Specifically, we use two key metrics from pupillometry: the Index of Pupillary Activity (IPA) and the Low/High Index of Pupillary Activity (LHIPA). We also include eye movement features, such as fixation gaze, blink behavior, and saccade behavior. We combined these features together and called them as Index Features. We further introduce the Diameter Series (DS) feature, which is generated by applying an overlapping time window to the diameter data. We hope DS features can capture the temporal dynamics and intricate patterns inherent in the data. It can be described as follows:

$$\mathrm{DS}(t, w) = \{D(t+i) \mid i = 0, 1, \ldots, w-1\}$$

where $\mathrm{DS}(t, w)$ represents the diameter measurement at time t, and w denotes the window size. w is the parameter we empirically set as 512 validated to have the best performance. By sliding the window across the data, the DS feature captures the evolving pattern of pupil diameter over time.

We evaluated the performance of well-established machine learning models in cognitive load classification tasks in HCI, including SVM [Bozkir et al. 2019; Szczepaniak et al. 2024; Yoshida et al. 2014], KNN [Bozkir et al. 2019], random forest [Bozkir et al. 2019; Szczepaniak et al. 2024; Yoshida et al. 2014], and XGBoost [Chen and Guestrin 2016; Souchet et al. 2022], as well as typical deep learning methods such as transformers and LSTM [Hochreiter and Schmidhuber 1997; Vaswani et al. 2017]. These models were tested across seven cognitive states, including three levels of working memory load, three levels of visual attention load, and a rest condition. While we explored deep learning approaches, their performance did not surpass that of traditional machine learning models and is therefore not detailed here. The machine learning algorithms were chosen for their representativeness of diverse classical approaches. SVM is a kernel-based method adept at handling high-dimensional data, while LDA, a statistical linear classifier, is particularly effective with Gaussian-distributed features. KNN offers a simple, non-parametric approach with local decision-making capabilities. RF, as an ensemble method, excels at managing noisy data and provides interpretable feature importance rankings. XGBoost, a gradient-boosting algorithm, is recognized for capturing complex feature interactions and delivering high performance on structured data tasks. Together, these algorithms provide a robust framework for evaluating the feature sets in this study. To ensure rigorous evaluation, we used an 80/20 split between training and testing data. Given the temporal nature of the data, we maintained a strict separation between training and testing periods to prevent data leakage. Models were trained on earlier time periods and tested on later periods, ensuring a reliable assessment of their generalization performance. As a foundational step in classifying cognitive load, we group it into two distinct types to simplify and focus the analysis.

According to Table 1, we found that index features demonstrated the best performance in distinguishing between working memory load and visual attention load. This result suggests that features derived from prior work possess a strong ability to differentiate these cognitive states and outperform the standalone DS features. It highlights the potential of additional feature engineering to further enhance the effectiveness of DS features. Moreover, SVM and LDA outperformed other machine learning models in this task, likely due to their ability to excel in scenarios where class boundaries are relatively well-defined. These results indicate that our formulation for distinguishing working memory load from visual attention load

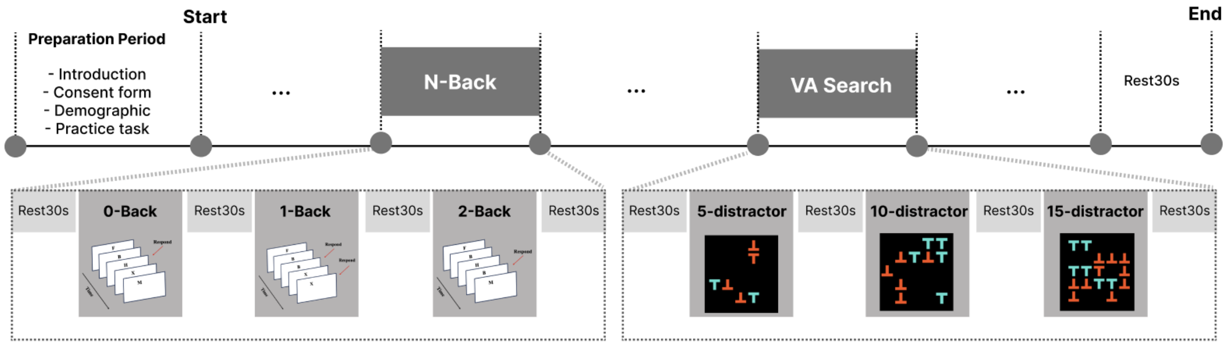

Figure 1: Experiment Design and Data Collection Procedure. The central timeline outlines the experimental sequence, starting with a preparation phase followed by eye tracker calibration. Participants then complete the main test, which includes two consecutive repetitions of each working memory task (0-back, 1-back, 2-back) and two consecutive repetitions of each visual attention task (5-distractor, 10-distractor, 15-distractor). The task order is counterbalanced to mitigate order effects. A 30-second rest period is included between tasks.

is feasible and effective. To further improve accuracy, future work should focus on advanced feature engineering, data augmentation techniques, and exploring hybrid models that combine the strengths of simpler classifiers like SVM or LDA with more complex ensemble methods.

Table 1: Performance Metrics for Models Across Features in Differentiating Working Memory Load vs. Visual Attention Load.

Feature	Model	Accuracy	Precision	Recall	F1 Score
Index Features	SVM	0.623	0.628	0.623	0.620
	LDA	0.622	0.626	0.622	0.620
	KNN	0.545	0.545	0.545	0.543
	RF	0.582	0.582	0.582	0.581
	XGBOOST	0.567	0.567	0.567	0.567
DS	SVM	0.552	0.554	0.552	0.550
	LDA	0.561	0.563	0.561	0.557
	KNN	0.528	0.529	0.528	0.523
	RF	0.540	0.548	0.540	0.523
	XGBOOST	0.548	0.552	0.548	0.540
Index + DS	SVM	0.600	0.600	0.600	0.600
	LDA	0.562	0.564	0.562	0.558
	KNN	0.545	0.545	0.545	0.543
	RF	0.564	0.567	0.564	0.559
	XGBOOST	0.593	0.608	0.593	0.579

5 DISCUSSION AND FUTURE WORK

This study advances the decoding of cognitive loads using eye-tracking data, providing a foundation for adaptive systems to better distinguish between working memory and visual attention. The strong performance of SVM and LDA underscores the well-defined class boundaries in this task, demonstrating the feasibility of fine-grained cognitive load classification. This suggests that the cognitive processes underlying working memory and visual attention manifest in distinct and measurable patterns that models may effectively capture.

In terms of feature extraction, we found the strong performance of index features, derived from prior research, indicates their robustness in capturing these distinctions. These features likely encode relevant temporal and spatial patterns that align closely with the cognitive mechanisms underpinning working memory and visual attention. In contrast, the underperformance of standalone DS features highlights their limitations in representing the nuanced characteristics of these cognitive states. This suggests a need for more advanced feature extraction techniques, such as deep learning-based embeddings, which could uncover richer representations of eye-tracking data by capturing non-linear and hierarchical patterns.

Moreover, individual differences may have played a significant role in the variability of model performance. Factors such as age, cognitive capacity, and prior experience can influence how individuals respond to tasks and exhibit cognitive load through eye-tracking metrics. For example, different people may show different pupil dynamics or fixation behaviors compared, potentially introducing bias into the dataset [Ha et al. 2021]. Future research should explore personalized approaches that account for these differences, such as individual calibration or stratified modeling techniques. Additionally, incorporating demographic and contextual information as supplementary features could improve the models' ability to generalize across diverse populations. Overall, addressing these challenges through advanced feature engineering, and personalized modeling could enhance the accuracy and applicability of cognitive load classification methods.

Finally, the modest overall accuracy highlights the complexity of fine-grained cognitive load classification. Cognitive load is influenced by various factors, including task design and environmental conditions, which can make it difficult to distinguish between subtle cognitive states. Our study provides a foundation for future research to develop more accurate classification techniques. While our focus was on differentiating cognitive load types, we did not explore the specific levels of each type, which could be a direction for future research. Building on these findings, future work can refine models and adaptive systems that dynamically respond to users' cognitive states, enabling more efficient, user-centered applications across a variety of fields.

6 CONCLUSION

This study introduces a first approach to differentiating between various cognitive load states, working memory vs. visual attention, using eye-tracking data. By leveraging multiple features, we demonstrated the feasibility to decode cognitive load in the HCI contexts. This approach has wide-ranging applications, from healthcare to education, where detailed cognitive load assessment can enhance performance and learning outcomes. Future research should explore further refinements of these techniques, applying them across more complex and dynamic environments.

REFERENCES

Davide Bacchin, Nina A. Gehrer, Krzysztof Krejtz, Andrew T. Duchowski, and Luciano Gamberini. 2023. Gaze-based Metrics of Cognitive Load in a Conjunctive Visual Memory Task. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 148, 8 pages. https://doi.org/10.1145/3544549.3585650

Alan Baddeley. 1992. Working memory. *Science* 255, 5044 (1992), 556–559. https://doi.org/10.1126/science.1736359

Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin* 91, 2 (1982), 276. https://doi.org/10.1037/0033-2909.91.2.276

Pradipta Biswas and Gowdham Prabhakar. 2018. Detecting drivers' cognitive load from saccadic intrusion. *Transportation research part F: traffic psychology and behaviour* 54 (2018), 63–78. https://doi.org/10.1016/j.trf.2018.02.002

Efe Bozkir, David Geisler, and Enkelejda Kasneci. 2019. Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)*. IEEE, 1834–1837. https://doi.org/10.1109/VR.2019.8797987

Federico Calà, Mustafa Can Gursesli, Pietro Tarchi, Lorenzo Frassineti, Andrea Guazzini, and Antonio Lanata. 2024. Pupil Dilation Dynamics in Serious Game Emotional Activation. In *2024 IEEE Gaming, Entertainment, and Media Conference (GEM)*. IEEE, 1–4. https://doi.org/10.1109/GEM51687.2024.00123

Siyuan Chen and Julien Epps. 2014. Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *Human–Computer Interaction* 29, 4 (2014), 390–413. https://doi.org/10.1080/07370024.2014.892428

Siyuan Chen, Julien Epps, Natalie Ruiz, and Fang Chen. 2011. Eye activity as a measure of human mental effort in HCI. In *Proceedings of the 16th international conference on Intelligent user interfaces*. 315–318. https://doi.org/10.1145/1943403.1943459

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 785–794. https://doi.org/10.1145/2939672.2939785

Aaron V Cicourel. 2004. Cognitive overload and communication in two healthcare settings. (2004).

Andrew T Duchowski, Krzysztof Krejtz, Nina A Gehrer, Tanya Bafna, and Per Bækgaard. 2020. The low/high index of pupillary activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3313831.3376632

Andrew T Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13. https://doi.org/10.1145/3173574.3173856

Stephen A Engel, David E Rumelhart, Brian A Wandell, Adrian T Lee, Gary H Glover, Eduardo-Jose Chichilnisky, Michael N Shadlen, et al. 1994. fMRI of human visual cortex. *Nature* 369, 6481 (1994), 525–525. https://doi.org/10.1038/369525a0

Marzieh Salehi Fadardi, Javad Salehi Fadardi, Monireh Mahjoob, and Hassan Doosti. 2022. Post-Saccadic eye movement indices under cognitive load: a path analysis to determine visual performance. *Journal of Ophthalmic & Vision Research* 17, 3 (2022), 397. https://doi.org/10.18502/jovr.v17i3.9467

Alan Gevins and Michael E Smith. 2003. Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical issues in ergonomics science* 4, 1-2 (2003), 113–131. https://doi.org/10.1080/14639220210159717

Amr Gomaa, Alexandra Alles, Elena Meiser, Lydia Helene Rupp, Marco Molz, and Guillermo Reyes. 2022. What's on your mind? A Mental and Perceptual Load Estimation Framework towards Adaptive In-vehicle Interaction while Driving. In *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Seoul, Republic of Korea) *(AutomotiveUI '22)*. Association for Computing Machinery, New York, NY, USA, 215–225. https://doi.org/10.1145/3543174.3546840

Stephen J Guastello, Anton Shircel, Matthew Malon, and Paul Timm. 2015. Individual differences in the experience of cognitive workload. *Theoretical Issues in Ergonomics Science* 16, 1 (2015), 20–52. https://doi.org/10.1080/1463922X.2014.958181

Unsoo Ha, Sohrab Madani, and Fadel Adib. 2021. WiStress: Contactless stress monitoring using wireless signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–37. https://doi.org/10.1145/3460423

Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909 arXiv:https://doi.org/10.1177/154193120605000909

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. oup Oxford.

Daniel Kahneman. 1973. Attention and effort.

Michael J. Kane, Andrew R. A. Conway, Taiko K. Miura, and Gregory J. H. Colflesh. 2007. Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33, 3 (2007), 615–622. https://doi.org/10.1037/0278-7393.33.3.615

Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W Woźniak. 2023. A survey on measuring cognitive workload in human-computer interaction. *Comput. Surveys* 55, 13s (2023), 1–39. https://doi.org/10.1145/3605737

Krzysztof Krejtz, Andrew T Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. 2018. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one* 13, 9 (2018), e0203629. https://doi.org/10.1371/journal.pone.0203629

Sandra P Marshall. 2002. The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th conference on Human Factors and Power Plants*. IEEE, 7–7. https://doi.org/10.1109/HFPP.2002.1042860

Ravi Naik, Alexandros Kogkas, Hutan Ashrafian, George Mylonas, and Ara Darzi. 2022. The measurement of cognitive workload in surgery using pupil metrics: a systematic review and narrative analysis. *Journal of Surgical Research* 280 (2022), 258–272.

Karl F. Van Orden, Wendy Limbert, Scott Makeig, and Tzyy-Ping Jung. 2001. Eye Activity Correlates of Workload during a Visuospatial Memory Task. *Human Factors* 43, 1 (2001), 111–121. https://doi.org/10.1518/001872001775992570 PMID: 11474756.

Elisa Perego, Colm Caffrey, Elisa Ghia, Henrik Gottlieb, Olli Philippe Lautenbacher, Siobhan Moran, Mauro Mosconi, Pilar Orero, Marco Porta, Keith Rayner, Elizabeth R. Schotter, Christopher John Taylor, and Bartholomäus Wissmath. 2012. Eye tracking in audiovisual translation. Aracne editrice, 83–104. https://doi.org/10.4399/97888548491364

Maud Ranchet, John Morgan, Abiodun Emmanuel Akinwuntan, and Hannes Devos. 2019. Exploring the cognitive workload during a visual search task in Parkinson's Disease. In *MDS 2019, International Congress of Parkinson's Disease and Movement Disorders*. 1–p.

Darrell S Rudmann, George W McConkie, and Xianjun Sam Zheng. 2003. Eyetracking in cognitive state detection for HCI. In *Proceedings of the 5th international conference on Multimodal interfaces*. 159–163. https://doi.org/10.1145/958432.958464

Mohammad R Saeedpour-Parizi, Shirin E Hassan, and John B Shea. 2020. Pupil diameter as a biomarker of effort in goal-directed gait. *Experimental Brain Research* 238, 11 (2020), 2615–2623. https://doi.org/10.1007/s00221-020-05888-0

Alexis D Souchet, Mamadou Lamarana Diallo, and Domitile Lourdeaux. 2022. Cognitive load classification with a stroop task in virtual reality based on physiological data. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 656–666. https://doi.org/10.1109/ISMAR55827.2022.00082

Gijsbert Stoet. 2010. PsyToolkit: A software package for programming psychological experiments and surveys. https://www.psytoolkit.org/ Accessed: 2024-09-10.

Maike Stoeve, Markus Wirth, Rosanna Farlock, André Antunovic, Victoria Müller, and Bjoern M Eskofier. 2022. Eye tracking-based stress classification of athletes in virtual reality. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5, 2 (2022), 1–17. https://doi.org/10.1145/3532720

John Ridley Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* 18, 6 (1935), 643–662. https://doi.org/10.1037/h0054651

John Sweller. 2010. Cognitive load theory: Recent theoretical advances. (2010). https://doi.org/10.1017/CBO9780511844744.004

John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational psychology review* 10 (1998), 251–296. https://doi.org/10.1023/A:1022193728205

Dominik Szczepaniak, Monika Harvey, and Fani Deligianni. 2024. Predictive Modelling of Cognitive Workload in VR: An Eye-Tracking Approach. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*. 1–3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 6000–6010. https://doi.org/10.5555/3295222.3295349

Qiuzhen Wang, Sa Yang, Manlu Liu, Zike Cao, and Qingguo Ma. 2014. An eye-tracking study of website complexity from cognitive load perspective. *Decision support systems* 62 (2014), 1–10. https://doi.org/10.1016/j.dss.2014.02.009

Christopher Wickens, William Helton, Justin Hollands, and Simon Banbury. 2021. *Engineering Psychology and Human Performance*. https://doi.org/10.4324/9781003177616

David L Woods, Mark M Kishiyama, E William Yund, Timothy J Herron, Ben Edwards, Oren Poliva, Robert F Hink, and Bruce Reed. 2011. Improving digit span assessment of short-term verbal memory. *Journal of clinical and experimental neuropsychology* 33, 1 (2011), 101–111. https://doi.org/10.1080/13803395.2010.493776

Yutaka Yoshida, Hayato Ohwada, Fumio Mizoguchi, and Hirotoshi Iwasaki. 2014. Classifying cognitive load and driving situation with machine learning. *International Journal of Machine Learning and Computing* 4, 3 (2014), 210.

Beste F Yuksel, Kurt B Oleson, Lane Harrison, Evan M Peck, Daniel Afergan, Remco Chang, and Robert JK Jacob. 2016. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5372–5384. https://doi.org/10.1145/2858036.2858381

Marco A Zenati, Kay B Leissner, Suzana Zorca, Lauren Kennedy-Metz, Steven J Yule, and Roger D Dias. 2019. First reported use of team cognitive workload for root cause analysis in cardiac surgery. In *Seminars in thoracic and cardiovascular surgery*, Vol. 31. Elsevier, 394–396.