Gaze-based Prediction of Cognitive Load in Augmented Reality

GREESHMA NERELLA, University of Texas at Dallas, USA
DAISY GAN, University of Texas at Dallas, USA
BRENDAN DAVID-JOHN, Virginia Tech, USA
RAWAN ALGHOFAILI, University of Texas at Dallas, USA

Cognitive load affects learning and task performance; specifically, increased cognitive load hinders an individual's ability to process information. In augmented reality (AR) interfaces, distracting notifications can also heighten cognitive load. Integrated gaze tracking offers a non-intrusive way to monitor cognitive states and provides the opportunity to predict and adapt to changes in cognitive load.

In this paper, we demonstrate how cognitive load prediction models can leverage built-in gaze-tracking data in AR to accurately predict cognitive load during search tasks. We collected gaze data from participants under both cognitively overloaded and non-overloaded states and analyzed gaze feature signatures to identify load-dependent patterns. We compared individual and group-trained models for predictive performance and generalizability. We initially used logistic regression, then tested tree-based ensemble models to improve performance. The best-performing XGBoost group model achieves a test AUC-ROC of 0.85. This work demonstrates robust cognitive load monitoring for AR tasks using built-in eye-tracking measurements.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Mixed / augmented reality**.

Additional Key Words and Phrases: Eye-Tracking, Cognitive Load, Augmented Reality, Behavioral Analysis

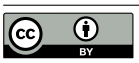
ACM Reference Format:
Greeshma Nerella, Daisy Gan, Brendan David-John, and Rawan Alghofaili. 2026. Gaze-based Prediction of Cognitive Load in Augmented Reality. *Proc. ACM Hum.-Comput. Interact.* 10, 3, Article ETRA026 (May 2026), 18 pages. https://doi.org/10.1145/3806040

1 Introduction

Reducing cognitive load is vital for optimizing human performance and learning [Tang et al. 2003; Tonnis et al. 2007; Yan et al. 2022]. This study is grounded in Cognitive Load Theory (CLT), which defines load as the demands placed on working memory during a task [Jeffri and Rambli 2021; Sweller et al. 1998]. CLT identifies three types of load: intrinsic (task complexity), germane (learning-related), and extraneous (environmental or design-related distractions) [Skulmowski and Xu 2022].

We specifically focus on extraneous load in immersive settings, which arises from processing non-task-relevant information like pop-up notifications [Rebello et al. 2024]. We treat these notifications as controlled manipulations of extraneous load, operationalized through observable behavioral and gaze-based responses to attention switching [Sweller et al. 2011; Thees et al. 2020]. Furthermore, we

Authors' Contact Information: Greeshma Nerella, Computer Science, University of Texas at Dallas, Richardson, Texas, USA, Greeshma.Nerella@UTDallas.edu; Daisy Gan, University of Texas at Dallas, Richardson, Texas, USA, Daisy.Gan@UTDallas.edu; Brendan David-John, Department of Computer Science, Virginia Tech, Blacksburg, Virginia, USA, bmdj@vt.edu; Rawan Alghofaili, Computer Science, University of Texas at Dallas, Richardson, Texas, USA, rawan@utdallas.edu.



ACM 2573-0142/2026/5-ARTETRA026
https://doi.org/10.1145/3806040

conceptualize load as a dynamic component of cognitive engagement that fluctuates as users respond to environmental demands [Booth et al. 2023]. Our method estimates how these interruptions influence mental effort and the overall immersive experience [Woodworth et al. 2023].

As Augmented Reality (AR) adoption grows, minimizing the extraneous load of notifications is essential, as alerts can degrade efficiency by diverting attention [Adamczyk and Bailey 2004; Bailey and Konstan 2006; Okoshi et al. 2015; Timileyin 2024]. Eye tracking offers a promising, non-intrusive method for monitoring these cognitive states [Katona 2022].

However, current research often relies on controlled desktop or virtual environments where participants remain stationary to ensure accuracy [Chen et al. 2011a,b; Duchowski et al. 2018; Jin et al. 2025]. These methods frequently fail in complex AR scenarios due to natural movement and visual noise. We bridge this ecological validity gap by collecting eye-tracking data during active AR tasks and labeling cognitive load based on real-time user responses to alerts.

In this paper, we evaluate the effectiveness of machine learning models in accurately classifying and predicting cognitive load in AR using eye-gaze data collected from participants completing tasks in an AR environment. We explore various predictor models, including ensemble methods, ultimately achieving a high degree of classification accuracy. The best-performing model achieved a test AUC-ROC of 0.85. Finally, we compare individual and group-trained models to gain insights into model generalizability for new users.

The major contributions of our work include:

- Demonstrate how specific eye-gaze features change (i.e., vary significantly) between conditions of cognitive overload and no cognitive overload.
- Validate an individualized machine learning model that utilizes eye-gaze data to accurately classify an individual's state as either experiencing cognitive overload or no cognitive overload.
- Assess the generalizability of a group machine learning model trained on aggregated eye-gaze data from multiple subjects to accurately classify the cognitive state (overload vs. no overload) across different individuals.

2 Related Works

In this section, we review the use of eye-gaze and other physiological data for measuring cognitive load in real-time. We specifically focus on how cognitive load prediction is a critical component of context-aware and everyday AR [Kim et al. 2025a].

2.1 Measuring Cognitive Load using Gaze Data

Eye-tracking provides a continuous, non-invasive method for assessing cognitive load. Recent studies have used it to decode cognitive states for personalized educational tools [Jin et al. 2025], building on historical methods using external cameras [Chen and Epps 2014; Rudmann et al. 2003] or wearable glasses [Biswas and Prabhakar 2018; Chen et al. 2011b] paired with VR or desktop screens [Bozkir et al. 2019]. While these systems accurately capture pupil diameter and fixation patterns [Li et al. 2024; Palinko et al. 2010], they typically require stable lighting and limited movement [Ktistakis et al. 2022; Van Orden et al. 2001].

Using built-in AR eye-tracking systems balances measurement accuracy with ecological validity [Chen et al. 2016]. While built-in tracking is common in VR headsets [Maquiling et al. 2024], we aim to expand the accessibility and scalability of these systems within AR [Zu et al. 2018]. Our work focuses on predictive modeling in representative environments where users move their heads freely [Awasthi et al. 2024; Caruso et al. 2021]. Machine learning is essential for processing the multidimensional gaze data required for real-time prediction [Ktistakis et al. 2022]. Robust feature

extraction is critical for state discrimination [Fuhl et al. 2023], with Random Forest classifiers being a common choice for this task [Kelleher and Hnin 2019; Oschlies-Strobel et al. 2017].

Prior research has successfully used Multi-layer Perceptrons and Random Forests to predict cognitive load scores validated by NASA-TLX in virtual training [Nasri et al. 2024]. Similar methods have identified specific gaze features, such as pupil area and movement, as significant predictors in 3D spaces [Gao et al. 2025]. While these methods often rely on high-fidelity VR or controlled settings [Khan et al. 2025], our approach extends this foundation to the dynamic constraints of AR [Mohanty et al. 2024].

2.2 Measuring Cognitive Load using Physiological Data

Beyond eye-tracking, biometrics such as heart rate variability, galvanic skin response, EEG, fMRI, and fNIRS provide precise, objective measures of cognitive load [Arjun et al. 2022; Chavarriaga et al. 2008; Ekin et al. 2025; Giannakos et al. 2019; Kalaganis et al. 2018; Lehne et al. 2009; Nourbakhsh et al. 2013; Wilbertz et al. 2018; Yuksel et al. 2016]. However, these technologies often require cumbersome equipment and controlled laboratory settings, creating a trade-off between accuracy and ecological validity [Darejeh et al. 2024].

In contrast, built-in AR eye-tracking offers a portable and ecologically valid alternative [Suzuki et al. 2023; Zagermann et al. 2016]. We propose modeling cognitive load using gaze data alone to simplify deployment and mitigate the sensor noise often associated with heart rate, EEG, and fNIRS in mobile settings.

2.3 Cognitive Load and Information Delivery

Cognitive load is a defining factor for effective information delivery in AR. Overlaid virtual content risks cognitive overload, which threatens user experience and safety by necessitating dual-tasking [Kim et al. 2025b]. Features such as ”always-on” guidance can become a burden, leading to attentional tunneling where users focus on virtual objects at the expense of their physical surroundings [Wickens and Alexander 2009]. To mitigate these risks, AR systems must be designed to actively reduce load [Tang et al. 2003; Yan et al. 2022]. Continuous assessment is essential for developing adaptive interfaces that optimize performance while maintaining real-world awareness.

3 Study

Established methods for inducing cognitive load utilize a dual-task paradigm to assess working memory demands, a core tenet of Cognitive Load Theory [Chen et al. 2016; Sweller et al. 2011]. In these setups, participants perform a primary task (e.g., problem-solving) alongside a secondary task (e.g., reacting to a stimulus). Performance degradation in the secondary task serves as a proxy for increased cognitive load in the primary task.

Similarly, our study employs a dual-task paradigm in immersive AR to investigate the link between interface interruptions and extraneous load. We combine a primary spatial color-matching game with a secondary ”load-alert” task to capture gaze-based data for machine learning analysis.

3.1 Participants

We recruited 31 college and graduate students with ages ranging from 18−30 (M=21 years, SD=1.96) to participate in our study. Nineteen of our participants identified as females, while twelve identified as males. Only one participant was completely unfamiliar with the concept of AR. Four of the participants have never used AR or VR headsets. All participants were screened to have normal vision. All participants provided written consent, and the study was approved by the university’s Institutional Review Board (IRB).

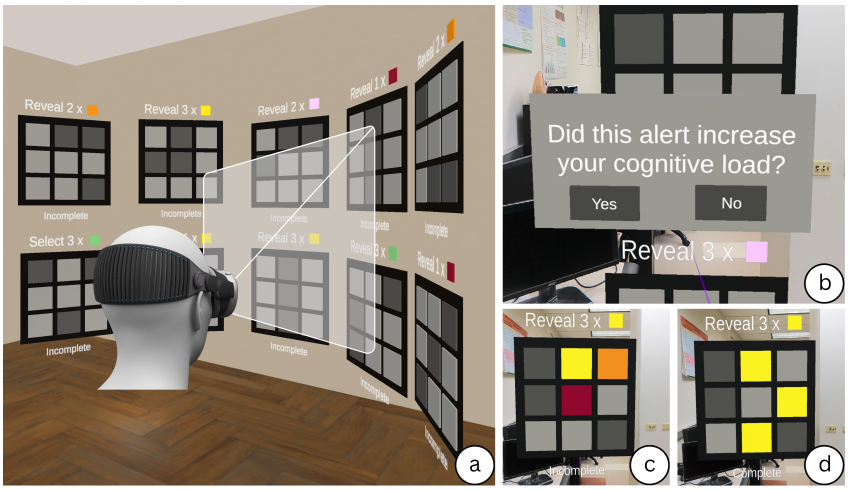


Fig. 1. Our user study environment and tasks. The full view of the ten boards in our AR game task is shown in (a). This view is not seen by the player. An example of a pop-up alert for our secondary task is shown in (b). (c-d) show the player view of one board with an instruction of "Reveal [3] × [yellow tiles]". (c) is an incomplete board, as there are remaining yellow colors to be revealed, as well as non-yellow colors to be unflipped. This is indicated by "Incomplete" at the bottom. (d) is the completed board after the participant revealed all three "yellow" tiles. This is indicated by "Complete" at the bottom.

3.2 Apparatus

Each participant wore a Magic Leap 2 Augmented Reality headset that tracks head orientation and eye gaze during the study. The headset utilizes cameras and infrared LEDs to estimate the location and orientation of the participant's eyes. We recorded gaze data using the Magic Leap 2 eye-tracking API at a frequency of 60 Hz. The study was implemented using Unity version 2022.3.11f1.

3.3 Study Design and Procedure

Upon arrival, participants reported demographics and received a briefing on cognitive load and AR cybersickness risks. After signing IRB consent forms, they fitted the Magic Leap headset and completed the standard eye-tracking calibration, which also served as an introduction to the controller.

We utilized a within-subject design consisting of five study blocks. In each block, participants played a five-minute AR color-matching game (Fig. 1), followed by a five-minute rest period and the NASA-TLX questionnaire [Hart and Staveland 1988]. This structure ensured consistent cognitive load assessment across the study.

The entire study, including preparation, orientation, and questionnaires, was conducted in person and lasted approximately $50 - 60$ minutes for each participant.

3.4 Tasks

We employed a dual-task paradigm to induce cognitive load akin to previous research [Chen et al. 2016; Sweller et al. 2011]. Each of our study blocks involved a primary decision task comprising of a spatial color-matching task and a simpler secondary distraction task for participants to indicate whether they are overloaded via an alert (Fig. 1).

In the primary color-matching task, we rendered ten boards each requesting the participant to reveal a number of tiles of a specific color (e.g., "Reveal [3] × [yellow tiles]"). The boards were arranged in a two-row five-column grid in an arc from -90° to 90° surrounding the participant's

head (Fig. 1a), forcing participants to physically turn their head and look up and down to view and interact with all ten boards. Each individual board presents a 3×3 grid of tiles with an instruction above (e.g., "Reveal [3] × [yellow tiles]"). Participants must flip a light gray tile in the board by selecting it to reveal its color. Three random tiles in each board are rendered in dark gray and disabled (i.e., cannot be flipped) to increase the game's difficulty. Only after the board is left with the instructed color tiles will the respective board's status below each board change from "Incomplete" (Fig. 1c) to "Complete" (Fig. 1d). Any additional color tiles or lack of instructed color tiles will still leave the board as "Incomplete". After completing all ten boards, the participant is assigned a new set of ten boards that replace the previously completed boards (i.e., a new color matching task). The tile colors in the boards as well as the instruction color are randomly assigned from a pool of 6 colors at every new assignment of ten boards. The instruction number is a randomly chosen integer from 1-3. All boards contain the required number of target-color tiles and are therefore solvable. To complete the primary color-matching tasks, the participant must solve the current set of ten boards (i.e., all ten boards must be marked as "Complete").

During the primary task, participants performed secondary distraction tasks by responding to alerts appearing at random nine-to-ten-second intervals. These alerts were positioned randomly along a 180° arc surrounding the participant and could appear both within and outside the field of view. Each alert asked, "Did this alert increase your cognitive load?" (Figure 1b), and participants were instructed to select "Yes" if they felt overloaded, but not to actively search for the alerts in game.

Each five-minute study block allowed participants to complete an unlimited number of color-matching tasks. Specifically, new sets of ten boards were presented until time expired. All interactions, including tile flipping and button presses, were performed using the Magic Leap controller's pointer and "Select" trigger.

3.5 Data Processing

Cleaning & Pre-processing. First, we cleaned and processed raw gaze data for consistency across participants. We identified points when the Magic Leap device may have failed to log data. In this scenario, large segments (> 1 s) of missing data were removed and smaller ($\leq$ 1 s) ones were interpolated using linear interpolation. To ensure consistency, gaze direction vectors are transformed from head-relative space to world space. Head angles are represented as quaternions (x, y, z, w). We convert these to rotation matrices that are then applied to the gaze direction vectors (x, y, z). The resulting eye-in-world vectors are then normalized to unit length to correct for rounding errors introduced during the rotation. This correction process ensures that gaze directions are recorded in a consistent world coordinate system for analysis.

Feature extraction. From the preprocessed gaze data, we extracted a total of 59 gaze features to capture aspects of eye movement. Gaze velocity was computed as the angular displacement between consecutive gaze vectors. The resulting velocity signal was smoothed using a seven-sample median filter to reduce noise for event detection. Saccade and fixation events were identified using standard velocity and dispersion-based algorithms. Saccades were detected using the Identification by Velocity Threshold (I-VT) with a velocity threshold of 70°/s. Fixations were identified using the Identification by Dispersion Threshold (I-DT) method with a dispersion threshold of 1°/s. [Salvucci and Goldberg 2000]. Using both detection methods to ensure robustness of fixation detection as each relies on distinct assumptions (velocity-based vs. dispersion-based) as seen in recent immersive eye-tracking studies [Peacock et al. 2022]. We are motivated by prior work showing that fixation metrics are sensitive to event detection methods and parameters [Shic et al. 2008]. Additionally, 3D Gaze signals in AR are affected by head motion, scene dynamics and other factors which can

lead to threshold sensitivity in a single detector. By evaluating both, we mitigate the risk that cognitive-load classification is heavily affected by the chosen fixation detection method.

Following event detection, we computed the features related to saccade dynamics and fixation behavior. These included measures such as fixation duration and saccade amplitude [Srivastava et al. 2018]. From these 59 features, we identified and selected the ten most informative features. We determined these features by calculating statistical significance using t-tests in Section 4.1 at each timestep in the maximum lens size. To prevent information leakage, this feature selection process was conducted exclusively on the training data so the test data remains unseen. We chose the ten features where the most number of time steps were statistically significant.

Windowing. We also analyzed the influence of lens size in predicting cognitive load, as can be seen in Section 4.2 and 4.3. These lenses were computed with a sliding window approach. These sliding windows are zero-padded at edges. Each window contained some $N = 3, 9, \ldots, 36$ timesteps and some $M = 0, \ldots, 59$ features. Each window captures the full temporal sequence of features $N \times M$, and did not use any summary statistics across the window.

Labeling. Each lens is labeled using a center-based approach to capture anticipatory gaze patterns. When a participant presses the "Yes" button in response to a distraction pop-up, that specific time step is marked with a positive label (i.e., Cognitive Overload). Lenses that have the overload time step at their center are designated as windows of cognitive overload, while all other lenses are labeled as having no cognitive overload.

Data. From the 31 participants, we recorded five blocks of five minutes (300 s) of gaze data. At 60Hz, this was 90,000 time steps per participant and 2,790,000 time steps in total (46,500 seconds). We split this dataset by participant to evaluate the model on unseen participant data, ensuring no overlap between the training and test sets. Individual and group model dataset splitting is discussed in Section 4.2. For the model comparison in Section 4.3, we extract 12% of this data to be used as our test set, while the remaining was used for training. Of this, we have 1,507 training samples with positive labels and 2,368,343 with negative labels, and rebalance the dataset to avoid overfitting on negative samples.

4 Results

Understanding how eye-gaze behavior reflects cognitive processes is essential for developing accurate cognitive load prediction systems. To get deeper insight into this relationship, we conducted a multi-level analysis examining gaze signatures of cognitive load. We also assessed the ability of both individual and group models to predict cognitive load. This approach allows us to uncover how gaze is linked to cognitive load, and evaluate model generalizability versus subject-specific adaptation. We used t-testing, the specific test details discussed separately for each hypothesis, the associated t and p values for each claim can be found in the supplementary materials. We explored three hypotheses related to measuring cognitive load using eye-gaze data similar to the work provided in this paper [Peacock et al. 2022]:

- **H1:** Gaze features vary under cognitive overload and no cognitive overload.
- **H2:** An individual model trained on eye-gaze can distinguish between states of cognitive overload and no overload.
- **H3:** A group model trained on eye-gaze can distinguish between states of cognitive overload and no overload.

4.1 Gaze Signatures of Cognitive Load

To examine the difference in gaze behavior between instances where participants experienced cognitive overload versus no cognitive overload, we visualized and compared the average feature value for both instances. The visualization spans the largest window size in our analyses (600 ms)

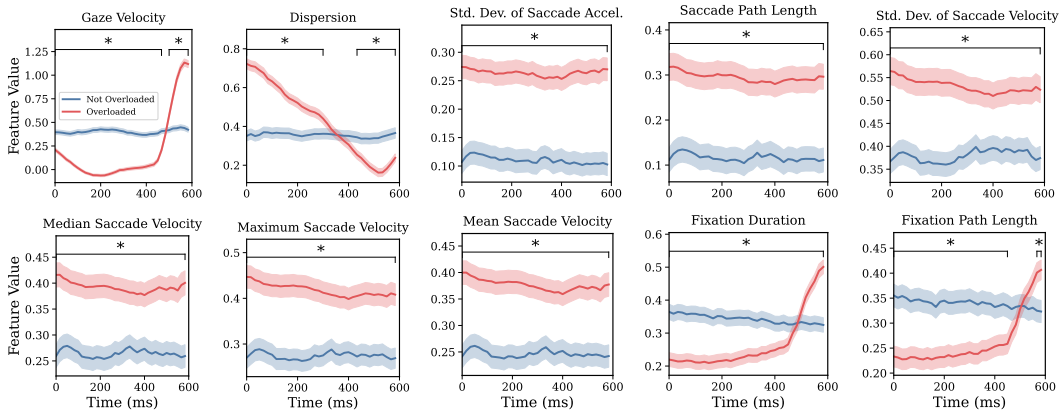


Fig. 2. Time-series plots of extracted gaze features.

[Peacock et al. 2022] to see the full temporal context in which gaze features may diverge. This 600 ms window produces a time series, a sequence of feature values ordered in time that allows us to observe how gaze behavior unfold between overload and no overload states. All feature values are normalized on a per-participant basis to scale feature values to be comparable across participants despite differences in baseline gaze characteristics. The resulting average feature trajectories, along with shaded regions representing the standard error, are shown as time-series plots in Fig. 2.

Then, we conducted a unpaired t-test throughout the values of the gaze feature between states of cognitive overload and no overload. This unpaired t-test revealed points of significant difference within these features between the overloaded and non-overloaded states. This supports the hypothesis that gaze features vary under cognitive overload and no cognitive overload (H1). These are indicated by the brackets above each plot and marked by an asterisk. We plot a time-series visualization of ten of these features in Fig. 2.

To identify which gaze features most clearly distinguished between cognitive overload and no overload states, we ranked all features based on the number of points in their time series that showed a statistically significant difference in the unpaired t-test. We choose to visualize the top ten features to show the most discriminative features and avoid redundancy since many of the next ranked features exhibited similar and weaker versions of the same patterns. This also proves to be useful feature selection for the logistic regression models trained in the following sections.

The pattern of feature values gives insight into difference in gaze behavior between cognitive overload and no-overload conditions. Saccade features are consistently elevated in cognitive load conditions, suggesting more rapid eye movements. Fixation features rise towards the end of the time series, possibly reflecting an attempt to refocus after an overload event. Gaze velocity shows a sharp increase during overload, while dispersion sharply decreases, indicating that gaze rapidly scans a smaller area. This behavior is consistent with recovery in the study, participants reorient to the specific area of the board they were working on, scanning only the relevant area to continue achieving the objective.

4.2 Comparing Individual and Group Models in Classifying Cognitive Load

To understand whether cognitive load can be measured from eye-gaze at both personal and general levels, we evaluate individual models trained separately for each participant. These models allow us to examine the extent to which cognitive load reflects participant-specific patterns. We compare

Table 1. AUC-ROC scores across individual and group models for the selected gaze features.

Feature	Individual Models		Group Models	
	AUC-ROC (50 ms)	Best AUC-ROC (Lens Size)	AUC-ROC (50 ms)	Best AUC-ROC (Lens Size)
All Features	.61	.74, 600 ms	.67	.76, 600 ms
Dispersion	.51	.65, 500 ms	.55	.67, 600 ms
Gaze Velocity	.64	.80, 600 ms	.62	.76, 600 ms
Std. Dev. of Saccade Velocity	.53	.53, 600 ms	.53	.54, 500 ms
Mean Saccade Velocity	.52	.53, 550 ms	.53	.54, 600 ms
Median Saccade Velocity	.52	.53, 550 ms	.53	.54, 600 ms
Maximum Saccade Velocity	.52	.52, 550 ms	.54	.54, 550 ms
Fixation Path Length	.53	.53, 100 ms	.54	.59, 600 ms
Fixation Duration	.53	.58, 600 ms	.53	.61, 600 ms
Saccade Path Length	.51	.53, 350 ms	.53	.54, 550 ms
Std. Dev. of Saccade Accel.	.51	.54, 500 ms	.54	.55, 600 ms

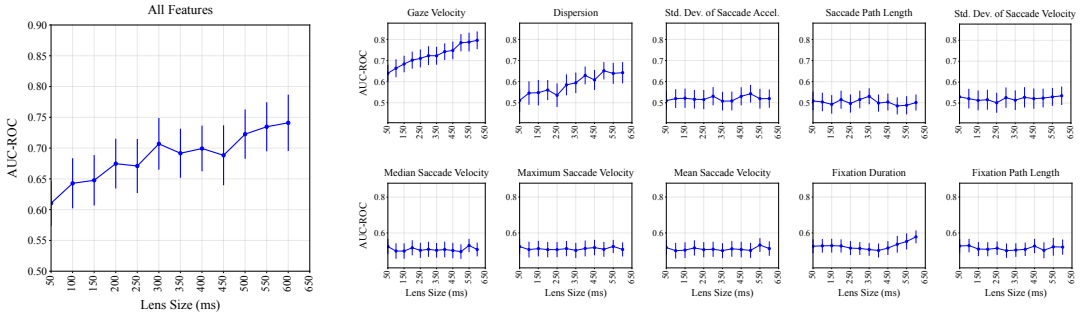


Fig. 3. Average AUC-ROC scores for individual models trained on all 10 features and individual features. Error bars represent 95% confidence intervals.

individual models to group-trained models and assess the generalizability of cognitive load measurement. A summary of both approaches is provided in Table 1. These models are trained on the best features determined by the largest number of significant timesteps earlier in Section 4.1.

Individual Models. We first explore the training of individual models personalized for each participant to predict cognitive load (Fig. 3). To accomplish this, we trained logistic regression models using four randomly sampled blocks for each participant and tested them on the remaining block. We conducted this analysis for various lens sizes and plotted the mean AUC-ROC score across participants in Fig. 3.

We conducted a one-sample one-tailed t-test with False Discovery Rate (FDR) correction to investigate the difference between the individual models and random chance prediction (0.5 AUC-ROC). For the model trained on all features, we successfully distinguished events of cognitive overload at all lens sizes, demonstrating consistent performance that was better than random chance ($p < .05$). The best model, with a lens size of 600 ms, achieved an AUC-ROC score of 0.74. By training the models on gaze velocity alone, we achieved an AUC-ROC score of 0.8, which was better than the score achieved by models trained using all features at lens sizes longer than 400 ms. We verify this using a one-tailed independent t-test, which results in a significant difference starting from a lens size of 400 ms and onward ($t = 2.14, df = 30, p < .05$). Classifiers trained on gaze velocity alone are perform significantly better than random at all lens sizes ($p < .05$). Dispersion becomes informative at later window sizes, reliably distinguishing events of cognitive load for lens sizes larger than 300 ms ($t = 3.4, df = 30, p < .05$). The performance of gaze velocity and dispersion models calls back to Section 4.1 where we can notice more varied signatures between the

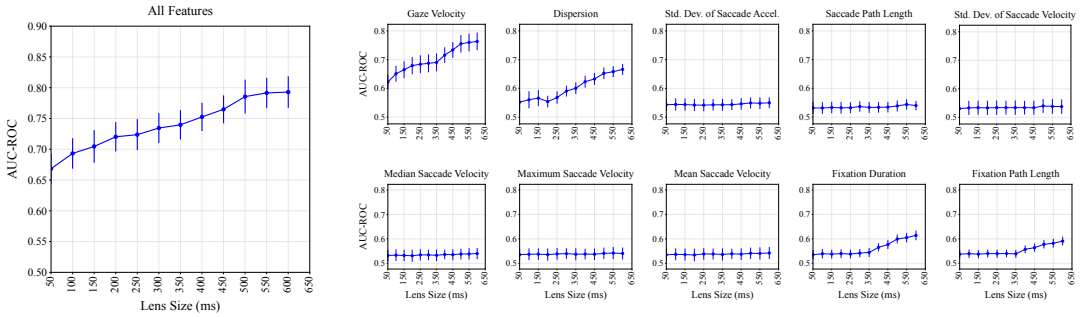


Fig. 4. Average AUC-ROC Scores for group models trained on all 10 features and individual features. Error bars represent 95% confidence intervals.

overload and no overload conditions for the two. Notably, with a lens size of 550 ms the models can distinguish cognitive load when trained with only gaze velocity, dispersion, fixation duration, and mean saccade velocity better than random ($p < .05$). Overall, these results support the hypothesis that an individual model trained on gaze data can distinguish between overload and no overload (H2).

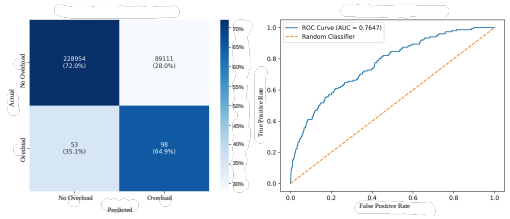
Group Models. Models trained on a large dataset that generalizes well across participants are often more practical as removing the requirement for per-user samples. We trained models using leave-one-subject-out iterations and plotted their average AUC-ROC score by varying the lens sizes (Fig. 4). In this evaluation, the ten most informative features identified in Section 4.1 were globally selected and kept consistent across all iterations. We achieved this by training a logistic regression classifier on the remaining data, after excluding a different participant each time for all lens sizes.

For a group model trained on all ten features, it is reliably able to distinguish states of cognitive overload and no cognitive overload at all lens sizes. We verify that the model performs significantly better than random using a one-sample t-tests ($p < .05$). Unlike the individual models, all models trained on one of the ten features perform statistically better than chance ($p < .05$). Notably, the models trained on gaze velocity alone performed just as well as the model trained on all features (t=-0.13, df=30, p > .05 at 600 ms). Combined, these results also support the hypothesis that a group model trained on eye-gaze can distinguish between states of cognitive overload and no cognitive overload (H3).

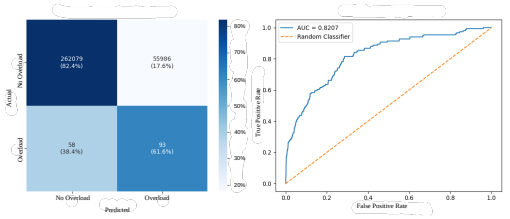
Comparison of Group and Individual Models. Individual models can identify eye-gaze patterns that indicate cognitive load for each participant. However, there are significant drawbacks to this approach, the most obvious being the requirement for extensive individualized data for effective model training. In this context, group models offer a more practical alternative if they can effectively detect instances of cognitive overload just as well as individual models.

We conducted unpaired t-tests to compare the differences between the AUC-ROC of the individual and group models trained on all the selected features at varying lens sizes (Fig. 6). These tests revealed a significantly better performance of group models over individual models for all lens sizes ($p < .05$) with the exception of 200 ms and 300 ms ($p = 0.0521$, $p = 0.25$).

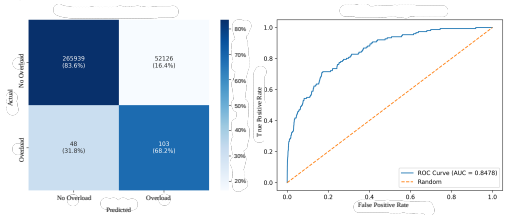
The difference in performance between group models and individual models in this experiment highlights the biggest drawback of personalized models: the need for a considerable amount of ground-truth data from each novel user. The personalized models had access to data from only four study blocks, with the remaining block reserved for testing. This resulted in an average of 1,450 seconds of data for training and 500 seconds for testing across participants. In contrast, group models had a larger amount of data from 30 participants completing five study blocks, with the remaining one participant's data held out for testing. Group model training data totaled an average



(a) Logistic Regression



(b) Random Forest



(c) XGBoost

Fig. 5. Confusion Matrix and ROC Curves for each group model type explored at a lens size of 550 ms

of 45,000 seconds across participants and lens sizes, while an average of 1,500 seconds was used for the test set. With more data from each user, individual models could potentially learn user-specific gaze patterns better than group models. However, considering the need for large amounts of data from each user, group models present themselves as a more practical option for developing systems based on or around cognitive load prediction, especially if these models can be later fine-tuned with small amounts of individual data to recognize the patterns from each new user [Masko 2017].

4.3 Experimenting with Other Machine Learning Classifiers

Extrapolating from H2 and H3, we wanted to explore how other model representations perform in distinguishing cognitive overload using eye-gaze data. After a preliminary search to find model types to experiment with, the following two were chosen: Random Forest and Extreme Gradient Boosting (XGBoost). Logistic Regression serves as the baseline model estimator in this set of experiments. Both XGBoost and Random Forest, while having different approaches, are tree-based ensemble models. These models have better inherent feature ranking due to their hierarchical structure and can use important features to arrive at better predictions of cognitive load. In other words, they enable us to investigate gaze features of cognitive load in greater detail.

Table 2. Performance comparison of group classification models.

Model	AUC-ROC	Precision	Recall	F1 Score
Logistic Regression	0.76	0.84	0.99	0.72
Random Forest	0.82	0.99	0.83	0.90
XGBoost	0.85	0.91	0.99	0.84

All of the following explorations were conducted with a lens size of 550 ms where the most number of significant features for individual and group models are seen in Section 4.2. This resulted in 2,369,850 lenses for training and 318,216 lenses for testing. These models are trained on all 10 of the previously explored features in Section 4.1 rather than individual features. Also, because we split our data by participant prior to extracting the training and test set (i.e., no overlap between the training and test sets), we can more reliably evaluate the model on unseen participant data.

Table 2 shows a summary of the performance of the logistic regression baseline, random forest, and XGBoost models. We present the class-weighted F1, precision, and recall scores to account for the imbalanced classes in our test data. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which provides a threshold-independent measure of discrimination ability, though it may be less sensitive to minority class performance in imbalanced datasets.

4.3.1 Logistic Regression. We first experimented with a Logistic Regression model to serve as our baseline. Unlike in our previous experiments with logistic regression, we tuned the hyperparameters using the Optuna framework to help the model achieve the maximize accuracy [Akiba et al. 2019]. We chose this hyperparameter tuning method due to its ability to improve the search space and tune continuous variables (e.g., inverse of regularization strength) with greater granularity by adaptively focusing on promising regions (TPE), early-pruning poor trials, and re-using knowledge. Given the size of the data, we excluded the Lib Linear solver in the tuning due to its slower convergence rate on large datasets. Among the remaining solvers, we varied regularization penalties and learning rates to identify the optimal configuration. We fitted a total of 100 estimators to the data using 5-fold cross-validation to tune these parameters.

The optimal model resulting from hyperparameter tuning used the SAGA solver [Defazio et al. 2014] with L_1 regularization and an inverse of regularization strength value of 0.38. This estimator achieved a mean validation AUC-ROC of 0.80 across the five folds. When evaluated on the test set, it yielded an AUC-ROC of 0.76, with about 72% of negatives and 64.9% of positives classified correctly (Fig. 5 (a)).

4.3.2 Random Forest. We used Random Forest to investigate the predictive performance of non-linear, tree-based models on the dataset. Another important feature of this model is that it ignores irrelevant features due to the random subsets on which each tree is trained. This reduces the need for particular feature selection to optimize performance in comparison to other models. We conducted hyperparameter tuning using Optuna as we did for logistic regression.

This resulted in a total of 50 trials, each building a separate random forest model to find the best estimator using 5-fold cross-validation. The best performing model of the 50 had 500 estimators, with unrestricted depths, features, and bootstrapping enabled. The chosen estimator had a 0.85 mean validation AUC across the five folds. Evaluation on the test set resulted in an AUC-ROC of 0.82 (Fig. 5 (b)). The model improves noticeably on logistic regression in predicting negatives (80% of total negatives predicted correctly), but exhibits a minor trade-off by missing 3% more positives.

4.3.3 XGBoost. For our final comparison, we evaluated the XGBoost classifier, which is another tree-based algorithm. One of the key advantages of XGBoost is its built-in feature selection process. As the model learns, each subsequent decision tree focuses more on the most effective features, while less informative features are used less frequently. As a result, the informative features tend to dominate the ensemble.

We tuned the hyperparameters of this model using Optuna for a total of 50 trials, each trial building a model. The best configuration of the 50 achieved a 0.88 mean validation AUC-ROC across 5-fold cross-validation with 728 estimators and a maximum tree depth of 7 (Fig. 5 (c)). This group model trained on the ten selected features from Section 4.1 resulted in the best performing model achieving the highest test set AUC among the three models of 0.85 AUC-ROC. The model is slightly better than Random Forest at predicting negatives. It is 4% more accurate at predicting positives than the Logistic Regression model.

Fig. 7 shows the top ten features by aggregated split frequency (weight), the number of times each feature was used in a tree split across all boosting rounds. In this context, we refer to it as aggregated feature importance because it consolidates the importance score of each feature across

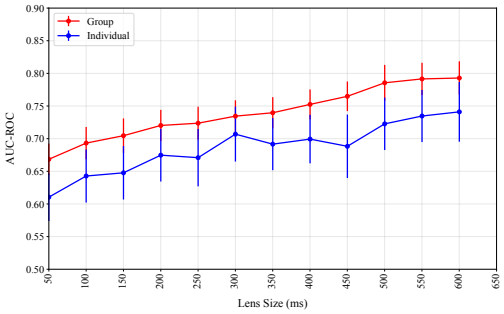


Fig. 6. Individual and Group models trained on all features.

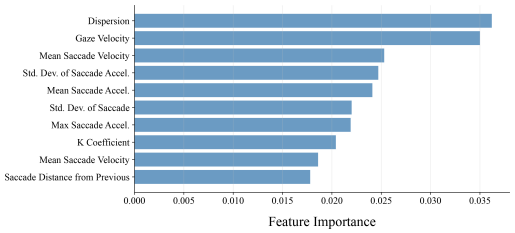


Fig. 7. Top ten features of our optimal XGBoost model.

all timesteps in the given lens size (i.e., 550 ms). We can see that the best performing features are still gaze velocity and dispersion akin to our earlier investigation of individual and group models in Section 4.2. Features that are more related to gaze movement speed (e.g., saccade velocity and acceleration) were prioritized by the XGBoost model over fixation features. This ranking further solidifies the role of gaze velocity for measuring overload using gaze data.

4.4 Discussion

Our exploration of gaze feature signatures reveals empirical evidence that these signatures vary across different cognitive load states (H1). When participants experience cognitive overload, their gaze movements become more rapid and erratic, as reflected in higher saccade feature values (Fig. 2). Fixation differences are subtle in the first 500 ms but increase in the last 100 ms, suggesting participants attempt to re-focus on the task. As a result, we observe that gaze tends to stabilize and exhibit longer fixations when participants are not under cognitive overload, accompanied by overall less rapid saccade movements. This pattern is particularly evident in the K coefficient, a validated measure of fixation stability indicating longer and more stable gaze patterns, which appears among the top ten most informative features identified by our XGBoost model [Krejtz et al. 2016]. This conclusion is further supported by the top ten most informative features identified by our XGBoost models (Fig. 7), which are mostly related to saccade features. Of the best performing features, gaze velocity and dispersion show the best individual performance with the remaining features performing at or slightly above random (0.5-0.6 AUC-ROC). However, group models including all ten features outperformed models trained with any individual feature like gaze velocity (0.79 AUC-ROC vs. 0.74 AUC-ROC) (Fig. 4). This points to a possibility that while the individual feature's ability to recognize events of cognitive load may be limited, combining feature signals introduces complementary information that better captures eye-gaze patterns indicating cognitive load. Gaze velocity and dispersion likely perform better since they directly capture temporal shifts and spatial trends without the noise of intermediate event detection, allowing for robust generalization.

Existing approaches in gaze-based cognitive load prediction achieved high accuracy using tree-based models and neural networks [Nasri et al. 2024]. While that study evaluate cognitive load in VR, our approach achieves an AUC-ROC of 0.85 which is comparable to the results achieved in established VR benchmarks (0.8-0.9 AUC-ROC), demonstrating that gaze-based signals are discriminative in AR despite being a less controlled environment.

Comparing individual and group models highlights the importance of sufficient data for effective personalization. Individual models often underfit with limited per-user samples. Notably, classifiers trained on gaze velocity or dispersion alone sometimes perform as well as models using all ten

features. Group models, though less personalized, benefit from larger, diverse datasets, learning more robust, generalizable gaze patterns and often outperforming individualized models when user-specific data is scarce.

In real-time AR systems, group models are a more practical choice, providing reliable performance without per-user calibration and reducing onboarding time. Individual models may still add value for long-term users, enabling fine-grained optimization. A practical deployment could start with a group model for new users, gradually tuned with personalized data. This personalization can go beyond accuracy, adapting to which types of prediction errors (false positive vs. false negative) are more costly for each user.

Tree-based ensemble methods, like random forests and XGBoost, consistently outperform logistic regression by capturing nonlinear relationships and hierarchical feature interactions. XGBoost shows the strongest performance, building trees sequentially to correct errors from previous ones. This iterative refinement makes it more expressive than the parallel averaging used in random forests, making sequential correction a more effective learning strategy.

Our overall findings highlight the feasibility and potential of measuring cognitive load specifically for real-world AR applications. This capability enables exploration and evaluation of how AR designs can reduce cognitive load.

5 Limitations and Future Work

Interpreting NASA-TLX Results. Our study repeated highly similar trials where the only variation was notification frequency. Because we did not explicitly manipulate high and low load conditions, participants experienced relatively uniform demands. Consequently, NASA-TLX scores showed no significant correlation with the frequency of high-load reports ($r = -0.2$, $p > 0.5$), as subjective judgments converged without distinct condition-driven anchors. This discrepancy suggests an evaluation gap between global subjective reflection and real-time cognitive state fluctuations.

Learning Curves. To explore how increasing training data affects model convergence, we attempted to plot learning curves. While this analysis was informative for group models, where sufficient data enabled meaningful iterative augmentation, it proved uninformative for individual models due to minimal augmentation opportunities (only 4 increments per user). Consequently, we excluded this analysis from the results section. Future work could adopt alternative study designs that generate sufficient per-user data to enable a more thorough exploration of individual model learning curves.

Integrating Physiological Measurements of Cognitive Overload. Previous research has shown the potential of integrating different types of data, such as brain activity and gaze behavior, to predict cognitive load in VR [Arjun et al. 2022]. These multimodal models provide more accurate predictions of cognitive load. Additionally, data obtained from an EEG device can provide a reliable reference point for labeling a user's gaze signal based on their cognitive state indicated by the EEG readings.

Challenges in Real-Time Prediction. Future research is necessary to assess the practicality of using these models in real-world AR applications and to examine how computational resources affect their performance. We have not yet evaluated the effectiveness of our models for real-time predictions. Therefore, we are uncertain about the potential latency when these models are implemented in AR applications. Our work focused on simpler ML models that have a higher likelihood to perform inference in real-time compared to deep neural networks [Amadori et al. 2022; Sarkar et al. 2019; Zhu et al. 2025] to specifically address this concern. Future work could measure inference times across AR headsets and present performance trade-offs across model hyperparameters.

Generalization and Ecological Validity. While our initial results are promising, the seated position of participants may limit the generalizability and ecological validity of the study across all AR activities. Future work aims to evaluate model performance across more diverse AR tasks to bridge this gap. Previous research [Alghofaili et al. 2019] successfully validated classifiers across different navigation scenarios, and GazeIntent [Narkar et al. 2024] showed that gaze-based models perform best on tasks with similar characteristics. However, systematically evaluating cross-task generalization by training on multi-task datasets remains a clear direction for future research.

Context-aware AR Interfaces. Implementing real-time cognitive load assessment has the potential to lead to the development of adaptive AR interfaces that can tailor information delivery to enhance user performance. Continuous assessment of cognitive load opens up opportunities for creating adaptive interfaces that are aware of the user's cognitive state [Jin et al. 2025]. Previous research has demonstrated the influence of information delivery on cognitive load [Okoshi et al. 2015]. Optimizing the timing and frequency of notifications displayed on an AR interface can also significantly improve a user's performance [Syiem et al. 2021; Yu et al. 2022]. In the future, we aim to integrate cognitive load prediction models into adaptive AR interfaces that respond in real-time to fluctuations in a user's cognitive state.

Privacy and Ethics. The collection of detailed eye-gaze data via AR headsets raises significant privacy and ethical concerns, as these signals can reveal sensitive information regarding demographics, cognitive states, and medical conditions [Abdrabou et al. 2025]. To safeguard participants, all gaze data in this study were de-identified and stripped of personally identifiable information prior to analysis.

Beyond data handling, the downstream application of behavioral models presents risks; models detecting high cognitive load could be misused to exploit periods of user vulnerability [Ruocco et al. 2024]. As these technologies mature, it is essential to proactively establish safeguards, transparency mechanisms, and responsible-use guidelines to prevent potential manipulation.

6 Conclusion

In this paper, we presented a novel analysis of cognitive load specifically in augmented reality (AR) experiences. We examined the gaze features associated with cognitive load and investigated how to personalize predictive models for classifying cognitive load. Additionally, we compared the effectiveness of various machine learning models in differentiating between overload and non-overload conditions, achieving a best-performing group XGBoost model at a 550 ms lens size using all ten selected features with an AUC-ROC of 0.85. Our findings confirmed the practical applications of machine learning models in predicting cognitive load states within AR environments. This research lays the groundwork for future efforts in designing AR experiences and interfaces by integrating gaze-based cognitive load analytics into usability evaluations.

References

Yasmeen Abdrabou, Süleyman Özdel, Virmarie Maquiling, Efe Bozkir, and Enkelejda Kasneci. 2025. From Gaze to Data: Privacy and Societal Challenges of Using Eye-tracking Data to Inform GenAI Models. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications (ETRA '25)*. Association for Computing Machinery, New York, NY, USA, Article 109, 9 pages. doi:10.1145/3715669.3726788

Piotr D. Adamczyk and Brian P. Bailey. 2004. If not now, when?: 2004 Conference on Human Factors in Computing Systems - Proceedings, CHI 2004. 271–278. https://www.scopus.com/pages/publications/4544334265

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2623–2631. doi:10.1145/3292500.3330701

Rawan Alghofaili, Yasuhito Sawahata, Haikun Huang, Hsueh-Cheng Wang, Takaaki Shiratori, and Lap-Fai Yu. 2019. Lost in Style: Gaze-driven Adaptive Aid for VR Navigation. In *Proceedings of the 2019 CHI Conference on Human*

Factors in Computing Systems (New York, NY, USA, 2019-05-02) *(CHI '19)*. Association for Computing Machinery, 1–12. doi:10.1145/3290605.3300578

Pierluigi Vito Amadori, Tobias Fischer, Ruohan Wang, and Yiannis Demiris. 2022. Predicting Secondary Task Performance: A Directly Actionable Metric for Cognitive Overload Detection. 14, 4 (2022), 1474–1485. doi:10.1109/TCDS.2021.3114162 Conference Name: IEEE Transactions on Cognitive and Developmental Systems.

Somnath Arjun, Archana Hebbar, Sanjana, and Pradipta Biswas. 2022. VR Cognitive Load Dashboard for Flight Simulator. In *2022 Symposium on Eye Tracking Research and Applications (ETRA '22)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3517031.3529777

Satyam Awasthi, Vivian Ress, Sydney Lim, Michael Beyeler, and Tobias Höllerer. 2024. Eye Tracking Performance in Mobile Mixed Reality. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (Orlando, FL, USA, 2024-03-16). IEEE, 1049–1050. doi:10.1109/VRW62533.2024.00321

Brian P. Bailey and Joseph A. Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. 22, 4 (2006), 685–708. doi:10.1016/j.chb.2005.12.009

Pradipta Biswas and Gowdham Prabhakar. 2018. Detecting drivers' cognitive load from saccadic intrusion. 54 (2018), 63–78. doi:10.1016/j.trf.2018.01.017

Brandon M. Booth, Nigel Bosch, and Sidney K. D'Mello. 2023. Engagement Detection and Its Applications in Learning: A Tutorial and Selective Review. *Proc. IEEE* 111, 10 (2023), 1398–1422. doi:10.1109/JPROC.2023.3309560

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)* (2019-03). 1834–1837. doi:10.1109/VR.2019.8797758 ISSN: 2642-5254.

Thomas J Caruso, Olivia Hess, Kenny Roy, Ellen Wang, Samuel Rodriguez, Coby Palivathukal, and Nick Haber. 2021. Integrated eye tracking on Magic Leap One during augmented reality medical simulation: a technical report. 7, 5 (2021), 431–434. doi:10.1136/bmjstel-2020-000782

Ricardo Chavarriaga, Pierre W. Ferrez, and José del R. Millán. 2008. To Err is Human: Learning from Error Potentials in Brain-Computer Interfaces. In *Advances in Cognitive Neurodynamics ICCN 2007* (Dordrecht, 2008), Rubin Wang, Enhua Shen, and Fanji Gu (Eds.). Springer Netherlands, 777–782. doi:10.1007/978-1-4020-8387-7_134

Fang. Chen, Jianlong. Zhou, Yang. Wang, Kun. Yu, Syed Z. Arshad, Ahmad. Khawaji, and Dan. Conway. 2016. Robust multimodal cognitive load measurement. In *Robust multimodal cognitive load measurement*. Springer International Publishing.

Siyuan Chen and Julien Epps. 2014. Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. 29, 4 (2014), 390–413. doi:10.1080/07370024.2014.892428 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07370024.2014.892428.

Siyuan Chen, Julien Epps, and Fang Chen. 2011a. A comparison of four methods for cognitive load measurement. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference (OzCHI '11)*. Association for Computing Machinery, New York, NY, USA, 76–79. doi:10.1145/2071536.2071547

Siyuan Chen, Julien Epps, Natalie Ruiz, and Fang Chen. 2011b. Eye activity as a measure of human mental effort in HCI. In *Proceedings of the 16th international conference on Intelligent user interfaces (IUI '11)*. Association for Computing Machinery, New York, NY, USA, 315–318. doi:10.1145/1943403.1943454

Ali Darejeh, Nadine Marcusa, Gelareh Mohammadi, and John Sweller. 2024. A critical analysis of cognitive load measurement methods for evaluating the usability of different types of interfaces: guidelines and framework for Human-Computer Interaction. arXiv:2402.11820 [cs] doi:10.48550/arXiv.2402.11820

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. 2014. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. arXiv:1407.0202 [cs.LG] https://arxiv.org/abs/1407.0202

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* (Montreal QC Canada, 2018-04-21). ACM, 1–13. doi:10.1145/3173574.3173856

Merve Ekin, Krzysztof Krejtz, Carlos Duarte, Andrew T. Duchowski, and Izabela Krejtz. 2025. Prediction of intrinsic and extraneous cognitive load with oculometric and biometric indicators. 15, 1 (2025), 5213. doi:10.1038/s41598-025-89336-y Publisher: Nature Publishing Group.

Wolfgang Fuhl, Anne Herrmann-Werner, and Kay Nieselt. 2023. A temporally quantized distribution of pupil diameters as a new feature for cognitive load classification. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications (ETRA '23)*. Association for Computing Machinery, New York, NY, USA, 1–2. doi:10.1145/3588015.3590116

Hong Gao, Yapeng Gao, and Enkelejda Kasneci. 2025. An Explainable Machine Learning Approach for Cognitive Load Detection in Virtual Reality Using Eye Tracking Data. In *Proceedings of the 2025 International Conference on Multimedia Retrieval* (Chicago, IL, USA) *(ICMR '25)*. Association for Computing Machinery, New York, NY, USA, 340–348. doi:10.1145/3731715.3733275

Michail N. Giannakos, Kshitij Sharma, Ilias O. Pappas, Vassilis Kostakos, and Eduardo Velloso. 2019. Multimodal data as a means to understand the learning experience. 48 (2019), 108–119. doi:10.1016/j.ijinfomgt.2019.02.003

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. doi:10.1016/S0166-4115(08)62386-9

Nor Farzana Syaza Jeffri and Dayang Rohaya Awang Rambli. 2021. A review of augmented reality systems and their effects on mental workload and task performance. 7, 3 (2021). doi:10.1016/j.heliyon.2021.e06277 Publisher: Elsevier.

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 *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications* (Tokyo Japan, 2025-05-26). ACM, 1–7. doi:10.1145/3715669.3725864

Fotis P. Kalaganis, Elisavet Chatzilari, Spiros Nikolopoulos, Ioannis Kompatsiaris, and Nikos A. Laskaris. 2018. An error-aware gaze-based keyboard by means of a hybrid BCI system. 8, 1 (2018), 13176. doi:10.1038/s41598-018-31425-2 Publisher: Nature Publishing Group.

Jozsef Katona. 2022. Measuring Cognition Load Using Eye-Tracking Parameters Based on Algorithm Description Tools. 22, 3 (2022), 912. doi:10.3390/s22030912

Caitlin Kelleher and Wint Hnin. 2019. Predicting Cognitive Load in Future Code Puzzles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300487

Rozemun Khan, Johannes Vernooij, Daniela Salvatori, and Beerend P. Hierck. 2025. Assessing Cognitive Load Using EEG and Eye-Tracking in 3-D Learning Environments: A Systematic Review. *Multimodal Technologies and Interaction* 9, 9 (2025). doi:10.3390/mti9090099

You-Jin Kim, Radha Kumaran, Jingjing Luo, Tom Bullock, Barry Giesbrecht, and Tobias Höllerer. 2025a. On the Go with AR: Attention to Virtual and Physical Targets While Varying Augmentation Density. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2025-04-25) *(CHI '25)*. Association for Computing Machinery, 1–16. doi:10.1145/3706598.3714289

You-Jin Kim, Radha Kumaran, Jingjing Luo, Tom Bullock, Barry Giesbrecht, and Tobias Höllerer. 2025b. On the Go with AR: Attention to Virtual and Physical Targets while Varying Augmentation Density. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan, 2025-04-26). ACM, 1–16. doi:10.1145/3706598.3714289

Krzysztof Krejtz, Andrew Duchowski, Izabela Krejtz, Agnieszka Szarkowska, and Agata Kopacz. 2016. Discerning Ambient/Focal Attention with Coefficient K. *ACM Trans. Appl. Percept.* 13, 3, Article 11 (May 2016), 20 pages. doi:10.1145/2896452

Emmanouil Ktistakis, Vasileios Skaramagkas, Dimitris Manousos, Nikolaos S. Tachos, Evanthia Tripoliti, Dimitrios I. Fotiadis, and Manolis Tsiknakis. 2022. COLET: A dataset for COgnitive workLoad estimation based on eye-tracking. 224 (2022), 106989. doi:10.1016/j.cmpb.2022.106989

Moritz Lehne, Klas Ihme, Anne-Marie Brouwer, Jan B.F. van Erp, and Thorsten O. Zander. 2009. Error-related EEG patterns during tactile human-machine interaction. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (2009-09). 1–9. doi:10.1109/ACII.2009.5349480 ISSN: 2156-8111.

Qingchuan Li, Yan Luximon, Jiaxin Zhang, and Yao Song. 2024. Measuring and classifying students' cognitive load in pen-based mobile learning using handwriting, touch gestural and eye-tracking data. 55, 2 (2024), 625–653. doi:10.1111/bjet.13394 _eprint: https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13394.

Virmarie Maquiling, Li Zhaoping, and Enkelejda Kasneci. 2024. Imperceptible Gaze Guidance Through Ocularity in Virtual Reality. arXiv:2412.09204 [cs] doi:10.48550/arXiv.2412.09204

David Masko. 2017. Calibration in Eye Tracking Using Transfer Learning. https://api.semanticscholar.org/CorpusID:43437839

Siddarth Mohanty, Jung Hyup Kim, Varun Pulipati, Fang Wang, Sara Mostowfi, Danielle Oprean, Yi Wang, and Kangwon Seo. 2024. Measuring Cognitive Workload in Augmented Reality Learning Environments Through Pupil Area Analysis. In *Augmented Cognition: 18th International Conference, AC 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part I* (Washington DC, USA). Springer-Verlag, Berlin, Heidelberg, 167–181. doi:10.1007/978-3-031-61569-6_11

Anish S Narkar, Jan J Michalak, Candace E Peacock, and Brendan David-John. 2024. GazeIntent: Adapting dwell-time selection in VR interaction with real-time intent modeling. *Proceedings of the ACM on Human-Computer Interaction* 8, ETRA (2024), 1–18.

Mahsa Nasri, Mehmet Kosa, Leanne Chukoskie, Mohsen Moghaddam, and Casper Harteveld. 2024. Exploring Eye Tracking to Detect Cognitive Load in Complex Virtual Reality Training. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 51–54. doi:10.1109/ISMAR-Adjunct64951.2024.00022

Nargess Nourbakhsh, Yang Wang, and Fang Chen. 2013. GSR and Blink Features for Cognitive Load Classification. In *Human-Computer Interaction – INTERACT 2013* (Berlin, Heidelberg, 2013), Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.). Springer, 159–166. doi:10.1007/978-3-642-40483-2_11

Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K. Dey, and Hideyuki Tokuda. 2015. Attelia: Reducing user's cognitive load due to interruptive notifications on smart phones. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (2015-03). 96–104. doi:10.1109/PERCOM.2015.7146515

Andreas Oschlies-Strobel, Sascha Gruss, Lucia Jerg-Bretzke, Steffen Walter, and Dilana Hazer-Rau. 2017. Preliminary classification of cognitive load states in a human machine interaction scenario. In *2017 International Conference on Companion Technology (ICCT)*. 1–5. doi:10.1109/COMPANION.2017.8287084

Oskar Palinko, Andrew L. Kun, Alexander Shyrokov, and Peter Heeman. 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. Association for Computing Machinery, New York, NY, USA, 141–144. doi:10.1145/1743666.1743701

Candace E. Peacock, Ben Lafreniere, Ting Zhang, Stephanie Santosa, Hrvoje Benko, and Tanya R. Jonker. 2022. Gaze as an Indicator of Input Recognition Errors. 6 (2022), 142:1–142:18. Issue ETRA. doi:10.1145/3530883

N. Sanjay Rebello, Jeremy Munsell, Prasanth Chandran, Lester Loschky, Yifeng Huang, Minh Hoai, and Sidney D'Mello. 2024. Mapping students' self-reported cognitive load, situational engagement, and attentional-cognitive states in an online multimedia learning module. 354–360. doi:10.1119/perc.2024.pr.Rebello

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* (Vancouver British Columbia Canada, 2003-11-05). ACM, 159–163. doi:10.1145/958432.958464

Martina Ruocco, Pejman Saeghe, Frederic Kerber, Jan Gugenheimer, Mark McGill, and Mohamed Khamis. 2024. From Redirected Navigation to Forced Attention: Uncovering Manipulative and Deceptive Designs in Augmented Reality through Retail Shopping . In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Computer Society, Los Alamitos, CA, USA, 720–729. doi:10.1109/ISMAR62088.2024.00087

Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, Florida, USA) *(ETRA '00)*. Association for Computing Machinery, New York, NY, USA, 71–78. doi:10.1145/355017.355028

Pritam Sarkar, Kyle Ross, Aaron J. Ruberto, Dirk Rodenburg, Paul Hungler, and Ali Etemad. 2019. Classification of Cognitive Load and Expertise for Adaptive Simulation using Deep Multitask Learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–7. doi:10.1109/ACII.2019.8925507 ISSN: 2156-8111.

Frederick Shic, Brian Scassellati, and Katarzyna Chawarska. 2008. The incomplete fixation measure. In *Proceedings of the 2008 symposium on Eye tracking research & applications*. 111–114.

Alexander Skulmowski and Kate Man Xu. 2022. Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. 34, 1 (2022), 171–196. doi:10.1007/s10648-021-09624-7 Place: Germany Publisher: Springer.

Namrata Srivastava, Joshua Newn, and Eduardo Velloso. 2018. Combining Low and Mid-Level Gaze Features for Desktop Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 189 (Dec. 2018), 27 pages. doi:10.1145/3287067

Yuko Suzuki, Fridolin Wild, and Eileen Scanlon. 2023. Measuring cognitive load in augmented reality with physiological methods: A systematic review. 40 (2023), 375–393. doi:10.1111/jcal.12882

John Sweller, Paul Ayres, and Slava Kalyuga. 2011. Measuring Cognitive Load. In *Cognitive Load Theory*, John Sweller, Paul Ayres, and Slava Kalyuga (Eds.). Springer, New York, NY, 71–85. doi:10.1007/978-1-4419-8126-4_6

John Sweller, Jeroen J. G. van Merrienboer, and Fred G. W. C. Paas. 1998. Cognitive Architecture and Instructional Design. 10, 3 (1998), 251–296. doi:10.1023/A:1022193728205

Brandon Victor Syiem, Ryan M. Kelly, Jorge Goncalves, Eduardo Velloso, and Tilman Dingler. 2021. Impact of Task on Attentional Tunneling in Handheld Augmented Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan, 2021-05-06). ACM, 1–14. doi:10.1145/3411764.3445580

Arthur Tang, Charles Owen, Frank Biocca, and Weimin Mou. 2003. Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2003-04-05) *(CHI '03)*. Association for Computing Machinery, 73–80. doi:10.1145/642611.642626

Michael Thees, Sebastian Kapp, Martin P. Strzys, Fabian Beil, Paul Lukowicz, and Jochen Kuhn. 2020. Effects of augmented reality on learning and cognitive load in university physics laboratory courses. 108 (2020), 106316. doi:10.1016/j.chb.2020.106316

Akintunde Timileyin. 2024. The Role of Cognitive Load in Shaping Web Usability Requirements. *Available at SSRN 5247018* (2024).

Marcus Tonnis, Christian Lange, and Gudrun Klinker. 2007. Visual Longitudinal and Lateral Driving Assistance in the Head-Up Display of Cars. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality* (2007-11). 91–94. doi:10.1109/ISMAR.2007.4538831

Karl F. Van Orden, Wendy Limbert, Scott Makeig, and Tzyy-Ping Jung. 2001. Eye Activity Correlates of Workload during a Visuospatial Memory Task. 43, 1 (2001), 111–121. doi:10.1518/001872001775992570 Publisher: SAGE Publications Inc.

Christopher Wickens and Amy Alexander. 2009. Attentional Tunneling and Task Management in Synthetic Vision Displays. 19 (2009), 182–199. doi:10.1080/10508410902766549

Gregor Wilbertz, Madhura Ketkar, Matthias Guggenmos, and Philipp Sterzer. 2018. Combined fMRI- and eye movement-based decoding of bistable plaid motion perception. 171 (2018), 190–198. doi:10.1016/j.neuroimage.2017.12.094

Jason W. Woodworth, Andrew Yoshimura, Nicholas G. Lipari, and Christoph W. Borst. 2023. Design and Evaluation of Visual Cues for Restoring and Guiding Visual Attention in Eye-Tracked VR. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 442–450. doi:10.1109/VRW58643.2023.00096

Zihan Yan, Yufei Wu, Yiyang Li, Yifei Shan, Xiangdong Li, and Preben Hansen. 2022. Design Eye-Tracking Augmented Reality Headset to Reduce Cognitive Load in Repetitive Parcel Scanning Task. 52, 4 (2022), 578–590. doi:10.1109/THMS.2022.3179954 Conference Name: IEEE Transactions on Human-Machine Systems.

Difeng Yu, Ruta Desai, Ting Zhang, Hrvoje Benko, Tanya R. Jonker, and Aakar Gupta. 2022. Optimizing the Timing of Intelligent Suggestion in Virtual Reality. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2022-10-28) *(UIST '22)*. Association for Computing Machinery, 1–20. doi:10.1145/3526113.3545632

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* (San Jose California USA, 2016-05-07). ACM, 5372–5384. doi:10.1145/2858036.2858388

Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. 2016. Measuring Cognitive Load using Eye Tracking Technology in Visual Computing. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (BELIV '16)*. Association for Computing Machinery, New York, NY, USA, 78–85. doi:10.1145/2993901.2993908

Gancheng Zhu, Xiaoting Duan, Zehao Huang, Rong Wang, Shuai Zhang, and Zhiguo Wang. 2025. GazeFollower: An open-source system for deep learning-based gaze tracking with web cameras. 8, 2 (2025), 1–18. doi:10.1145/3729410

Tianlong Zu, John Hutson, Lester C. Loschky, and N. Sanjay Rebello. 2018. Use of Eye-Tracking Technology to Investigate Cognitive Load Theory. arXiv:1803.02499 [physics] doi:10.48550/arXiv.1803.02499