SensCogAR: Cognitive Load Estimation Via Movement Data in Assembly Tasks

Javier Melo[1*], Leyla Akinci[2*], Ko Watanabe[3], Nicolas Großmann[4], Shoya Ishimaru[5], Andreas Dengel[6]
[12346]DFKI GmbH, Kaiserslautern, Germany
[5]Osaka Metropolitan University, Osaka, Japan

Abstract

Understanding cognitive load in Augmented Reality (AR) applications has become increasingly relevant, especially within assembly instruction systems. To effectively address this challenge, we designed an experiment employing a task that shares key cognitive and motor characteristics with industrial assembly processes and can be manipulated to induce low and high levels of cognitive load. Participants completed tangram puzzles across two sessions representing these load levels. We collected physiological and movement data from wearable devices, including the Microsoft HoloLens 2 and Empatica E4, alongside NASA Task Load Index (NASA–TLX) scores and task completion times. Machine learning models trained on movement data (head, hand, and eye tracking from the HoloLens) achieved the highest classification F1 score of 0.886, outperforming combined sensor data and physiological data alone. NASA–TLX scores and task completion times validated the experimental manipulation of cognitive load. Our findings provide evidence that movement data captured via AR headsets can effectively detect cognitive load, with implications for adaptive AR assembly systems.

1 Introduction

AR has gained significant attention for interactive assembly instructions, aiming to improve user experience compared to traditional methods [6, 12, 14]. However, the extent to which these improvements have been achieved remains uncertain, underscoring the need for rigorous evaluation.

[1]javier.melo@dfki.de
[2]bin98hah@rptu.de
[3]ko.watanabe@dfki.de
[4]nicolas.grossmann@dfki.de
[5]ishimaru@omu.ac.jp
[6]andreas.dengel@dfki.de
*These authors contributed equally to this research.

Evaluations typically rely on performance outcomes or user experience metrics [9, 36], predominantly Cognitive Load (CL) [31]. Accurate CL estimation is crucial, as high load correlates with undesired outcomes such as workplace accidents and reduced satisfaction [5, 13, 26], whereas optimized CL enhances motivation and engagement [2, 37, 38].

While self-reports like NASA-TLX [15] are common, they are subjective and fail to capture continuous cognitive fluctuations due to their retrospective nature [30]. Physiological measures allow continuous monitoring [1, 3, 27] but often require intrusive hardware. Conversely, movement data embedded in AR headsets offers a non-intrusive alternative [17]. While eye movements have been used to estimate cognitive states Majumdar et al. [19], Melo et al. [21], Tanaka et al. [34], the use of comprehensive body movements to estimate CL remains largely unexplored, presenting an opportunity for device-free estimation.

To reliably infer CL from physiological and movement data, we require a dataset correlating these measures with established CL conditions, a resource currently lacking for manual assembly contexts. To address this gap, we engaged participants in solving tangram puzzles, serving as proxy for scenarios where objects of comparable size require manual assembly. By varying puzzle complexity, we created two experimental conditions representing high and low CL. We validated the experimental manipulation using NASA–TLX-scores and completion times, while simultaneously recording multi-modal data. Specifically, we captured movement data with a Microsoft HoloLens 2 AR headset and physiological signals with an Empatica E4 wristband.

Our central research question explores whether CL can be accurately classified using movement data captured by an AR headset, leveraging machine learning techniques. The following are the two main primary contributions (C1–C2) of this study:

C1. **Introduce a resource-efficient approach for detecting CL using only movement data from an AR headset.** Our findings demonstrate that models trained on data from the integrated sensors of the HoloLens 2 can reliably classify cognitive load. This eliminates the need for additional physiological sensors, or a separate workstation for off-device processing, making the approach suitable for resource-constrained environments.

C2. **We make our AR headset and wristband dataset publicly available [7].** We provide a consistent and reusable resource linking physiological and movement data to high and low cognitive load, enabling future studies to evaluate CL during manual assembly tasks and facilitating the development of adaptive assistance systems.

2 Related Work

To contextualize our approach, we first review the literature on AR-based instructional systems and their evaluation in industrial settings, followed

[7] Public Dataset: `https://osf.io/45tjp`

Table 1: Comparison of CL estimation studies to AR-based assembly scenarios.

Study Reference	N	AR	Assmb.	Perf.	Subj.	Phys.	Movt.	Public
Liang et al. [18]	11	✗	✗	✗	✗	✗	✓	✗
Vanneste et al. [35]	46	✗	✓	✗	✗	✓	✗	✗
Wiedenmaier et al. [39]	36	✓	✓	✓	✗	✗	✗	✗
Hou et al. [16]	20	✓	✓	✓	✓	✗	✗	✗
Dorloh et al. [12]	21	✓	✓	✓	✓	✗	✗	✗
Funk and Schmidt [14]	16	✓	✓	✓	✓	✗	✗	✗
Hou et al. [17]	20	✓	✗	✓	✓	✓	✓	✓
Ours	21*	✓	✓	✓	✓	✓	✓	✓

Note. **N** = number of participants; **AR** = use of AR; **Assmb.** = participants performed an assembly task; **Perf.** = study reports performance measures, e.g., task completion time; **Subj.** = use of subjective self-report measures, e.g., NASA–TLX; **Phys.** = physiological measures were used to predict cognitive load, e.g., Electrodermal Activity (EDA). **Movt.** = movement measures were used to predict cognitive load, e.g., eye tracking; **Public** = public availability of the data. *N varied depending on the modalities used for training. $N = 21$ considers movement data only.

by research on integrating CL estimation methods into AR interfaces. Table 1 provides a comparison of key studies discussed in this section.

2.1 AR-based Instructional Systems

Early research comparing AR to paper instructions relied on task completion times, noting performance gaps linked to cognitive processes without directly quantifying them [39]. Later studies incorporated subjective measures like NASA–TLX or System Usability Scale (SUS) [12, 14, 16], yet these retrospective ratings miss fine-grained cognitive fluctuations.

2.2 Sensors for Cognitive Load Estimation

Physiological methods enable continuous monitoring of cognitive states through biometric signals, overcoming the temporal limitations of subjective reports [30]. Vanneste et al. [35] utilized EDA, Electroencephalography (EEG) and Electrooculography (EOG) to detect CL elicited by an assembly task presented in three levels of difficulty. Although they found significant differences in CL among the difficulty levels, they reported a small effect size ($R^2 = 22.8\%$), indicating that physiological data may not be sufficient. Furthermore, using physiological sensors in real-world AR settings introduces a significant trade-off regarding intrusiveness.

In AR contexts, movement-based measures gain relevance, as modern AR devices are equipped with built-in sensors that track hand, head, and eyes. While research has focused on individual modalities, such as eye tracking [19] or the influence of CL on hand selection Liang et al. [18], integrating these streams into a cohesive, non-intrusive model for CL classification remains under-explored. This research gap presents an opportunity to leverage full multi-modal data captured by AR headsets.

In this study, we categorize all data streams into two groups: physiological and movement data. Although movement is a physiological process, we treat it separately to compare these two sources directly.

2.3 Positioning of the Present Work

Hou et al. [17] explored CL estimation in a Mixed Reality (MR)-based simulated Computer Numerical Control (CNC) machine operation, where participants responded to instructions through button presses. Using movement data from the HoloLens 2, and heart rate via a dedicated wristband, they tested the effect of noise and distractions on the participants' workload, as measured by the NASA–TLX. The authors trained machine and deep learning models on participant data to classify between their 3 experimental conditions. They achieved a peak $F1 = .96$ with Transformer models trained on movement data, while adding Heart Rate (HR) data decreased performance to $F1 = .83$. Despite these scores, button-pressing tasks lack the sensorimotor coordination required for manual assembly, potentially limiting the generalizability of their movement-based classifiers to assembly contexts.

Additionally, the computational intensity of Transformers may exceed the resources of current AR headsets. Hou et al. [17] reported significantly lower performance ($F1 = .73$) with lighter Random Forest (RF) models. Methodologically, their standardization process across the whole dataset suggests potential data leakage, and the absence of a clear cross-validation strategy makes generalization to unseen users uncertain.

Our study addresses these gaps by evaluating CL classification within a manual assembly scenario. We conducted a controlled experiment and analyzed task outcomes using objective and subjective measures. We also collected physiological and movement data to train lightweight machine learning models to distinguish between high and low levels of CL. We implemented a rigorous, leak-free standardization and leave-one-participant-out cross-validation to ensure model robustness. Finally, we publicly provide our multi-modal dataset, collected in Germany, to promote future research.

3 Data Analysis

3.1 Preprocessing

Our data consisted of physiological measurements from the Empatica E4 wristband and movement data from the HoloLens 2. The sensors on both devices sampled the data at different frequencies. From the Empatica E4, we collected Blood Volume Pulse (BVP) at 64 Hz, EDA at 4 Hz, and skin temperature at 4 Hz. From the HoloLens 2, we collected data on head position, hand position and orientation, as well as eye movements, all of which were sampled at 30 Hz.

To work with all sensors simultaneously, we divided the data into sequential windows, each containing 30 seconds of time series data, with an overlap of 10 seconds. We applied transformations to the raw signals from each data window and then characterized these signals by extracting

a comprehensive set of statistical and spectral features. The statistical features included mean, range, standard deviation and kurtosis, while the spectral features included entropy and the total Power Spectral Density (PSD).

3.2 Feature Engineering

3.2.1 Physiological Data Processing

To process physiological data we used NeuroKit2, a Python package to process these type of data [20]. From the BVP signal, we extracted R-R Intervals (RRI), which measure the time between heartbeats. Features extracted from RRI reflect Heart Rate Variability (HRV), which has been found to correlate with CL [29]. We extracted the tonic and phasic components from EDA. As noted by Melo et al. [21], this decomposition provides better results when measuring CL, compared to using the raw signal alone. Skin temperature was left in its original form.

3.2.2 Movement Data Processing

We calculated the empirical sample rate from the HoloLens data, which differed from the theoretical sample rate of 30 Hz. To do so, we computed the average time interval between consecutive timestamps and took its inverse. The resulting empirical sample rate was $21.34Hz$, which we used for all subsequent feature calculations dependent on the sampling frequency.

Processing Eye Blinks Eye movement data consisted of 3D vector pairs representing origin (position between both eyes) and direction (a point in space in the same vector from the origin to the gaze point) over time. A recording was marked as "missing value" if the eye tracking embedded in the HoloLens 2 could not detect any of the eyes, due to the participant blinking or not looking through the visor. While we did not find any study characterizing "out of visor" gaze patterns, there are studies characterizing blinks, from which we determined time thresholds.

Based on Doane [11], we adopted a mean blink duration (μ) of $257.9ms$ and a Standard Error of the Mean (SEM) of $11.3ms$, calculated from an average of 40 blinks. We calculated the standard deviation of blink durations by multiplying the SEM by the square root of the number of blinks ($N = 40$), resulting in a standard deviation of $71.47ms$. Since the authors did not specify distribution parameters for blink duration, we employed a normal distribution. We calculated the interval containing 99% of the blinks to include as many valid blinks as possible. This resulted in a range from $62ms$ to $454ms$, which we used to filter consecutive missing gaze data, treating only clusters within this range as blinks. We ignored clusters outside this range, assuming these corresponded to situations where participants did not look through the visor. From the detected blinks, we calculated their durations and inter-blink intervals.

Processing Gaze Fixations Calculating fixations from the HoloLens 2 gaze data presented a unique challenge. The device provides a 3D gaze vector, but its method for determining depth is not based on binocular convergence. We discovered that the gaze vector three meters away from a wall had almost the same coordinates as a gaze vector focusing on an object hanging halfway to the same wall. This sensor-specific behavior made it difficult to directly adapt standard fixation detection algorithms with thresholds set for different devices and tasks. To our knowledge, no publicly available HoloLens 2 eye-tracking dataset with ground-truth fixation data for a complex motor task exists.

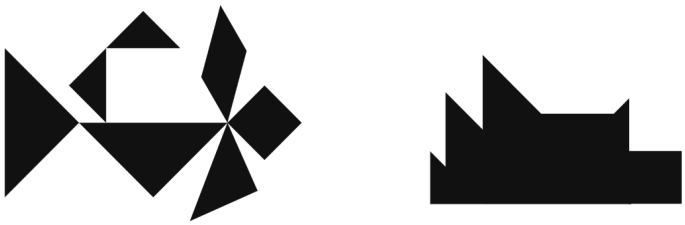
To address this, we implemented Identification by Dispersion-Threshold (IDT), a dispersion-based algorithm described in Salvucci and Goldberg [25]. The IDT algorithm considers consecutive gaze points as a fixation if they remain within a specified spatial distance and occur within a minimum time interval of $100ms$ [25]. Given the sampling rate of our data ($21.34Hz$), we set the minimum fixation threshold to 2 gaze points, which includes any fixation longer than $46.9ms$ to capture as many valid fixations as possible. While our threshold is lower than the $100ms$ recommended by Salvucci and Goldberg [25], we found it to be a good balance between capturing valid fixations and reducing false positives, as the next possible minimum threshold of 3 gaze points would have excluded fixations lower than $140.58ms$.

We extended the algorithm described by Salvucci and Goldberg [25] by adding a maximum dispersion threshold to account for gaze drift after initial fixation onset, following Buscher et al. [8]. To determine the dispersion thresholds for the algorithm, we analyzed the dataset provided by Aziz and Komogortsev [4], which employed a "follow the dots" task while recording eye-tracking data using a HoloLens 2 device, as in our study. The dataset provides a known quantity of visual stimuli (dots) that participants were instructed to fixate on, and we used it to optimize the dispersion thresholds for our algorithm.

We defined $47mm$ as the minimum fixation size and $95mm$ as the maximum fixation size, based on our empirical optimization of the thresholds against the Aziz and Komogortsev [4] data, such that the detected fixation count was as close as possible to the number of dots participants fixated on. From the fixations, we calculated duration, speed, acceleration and distance to previous fixation.

Processing Head Movements Head movements data collected from the HoloLens 2 consisted of position in 3D space, a normalized 3D vector indicating forward direction and a normalized 3D vector indicating upward direction. We calculated position change, and angular distance of the forward and upward vectors.

Processing Hand Movements Hand movements were recorded by the HoloLens 2, as long as the hands were in front of the participant. For each data point, each hand is represented by a position vector in 3D and an orientation vector in a 4D unitary quaternion. For each hand, we calculated angular distance (based on their orientation), position change.



(a) Low-complexity puzzle (b) High-complexity puzzle

Figure 1: Example puzzles used in the experiment. In low-complexity puzzles (left), all piece contours are visible, while in high-complexity puzzles (right), piece boundaries are harder to distinguish, increasing the difficulty of the task.

Additionally, based on the findings of Liang et al. [18], we calculated contralateral distance, which we operationalized as the distance from a hand along an axis that is perpendicular to a "midline", an imaginary forward-facing vertical plane that divides the body in two halves. We determined this midline by using the head position and forward vectors. Due to most of the participants being right handed, we calculated the maximum, mean and minimum of all these features across hands, so that the model would not be biased towards one hand over the other. Moreover, this decision mitigated the influence of missing data when at least one hand was outside the field of view of the hand tracker.

4 Data Collection

4.1 Participants

We recruited 24 participants, all students or employees in Germany. Due to technical issues, data from one participant were excluded, resulting in a final sample of 23 (15 female, 7 male, 1 non-binary). Most participants completed a bachelor's degree; four held master's degrees and one held a doctoral degree. All but one participant, who was left-handed, reported being right-handed. During data analysis, we discovered that additional data were missing due to connection problems with our devices. Physiological data were missing or incomplete for five participants, and movement data were missing for two participants. This led to the following sample sizes by modality: 21 participants for the movement-only dataset, 18 participants for the physiological-only dataset, and 17 for the dataset combining both physiological and movement data. All participants reported either no prior experience with tangram puzzles or having played only a few times more than six months prior. Before the experiment, all participants provided informed consent, and the study was approved by the ethics board (DFKI Ethics Board). When applicable, participants received one participation hour as course credit in accordance with university requirements.

4.2 Experimental Setup

We used Tangram puzzles as proxies for manual assembly tasks involving the manipulation of objects of similar size. Each puzzle consists of seven simple geometrical pieces that participants must assemble to match a reference image, where the location and orientation of each piece are obfuscated. One advantage of selecting Tangram as a proxy for assembly scenarios is that it allows for experimental manipulation of cognitive load, consistent with Cognitive Load Theory (CLT)[31, 32], as shown by Vanneste et al. [35].

We designed 47 tangram puzzles using the Tangram Builder tool [23], considering two difficulty levels, determined by the complexity of the puzzles. Following Vanneste et al. [35], puzzle difficulty was estimated by the visibility of piece contours: in low-difficulty puzzles, all contours were fully visible, while in high-difficulty puzzles, most contours were partially obscured. Example puzzles from each complexity level are shown in Figure 1.

Similarly, we grouped the tangram puzzles into two groups based on their complexity, to create two experimental conditions that would elicit distinct levels of cognitive load, which could be detected by differences in behavioral, physiological, movement and subjective outcomes. To ensure a sufficient number of trials across the experiment, 40 puzzles were designed to have a low complexity and seven to have a high complexity. This distribution was selected based on preliminary testing, that suggested that low-complexity puzzles could be solved in under a minute, while high-complexity puzzles required several minutes to be solved.

We developed an experiment using PsychoPy [22] to guide participants through the puzzles and to record events, such as trial start and end times. This allowed for behavioral analyses of the data. The experiment application was presented on a computer in front of the participants. This experimental setup, depicted in Figure 2, was chosen to mimic scenarios where participants assemble objects while seated, following instructions presented in front of them on a screen or through a head-mounted display.

Throughout the experiment, the participants wore two devices we used to collect data we then analyze to estimate their cognitive load. A Microsoft HoloLens 2 AR headset was worn to collect movement data, consisting of eye movements, head position, and hand position and orientation. Raw data collection was possible thanks to the HL2SS plugin [10]. The participants also wore an Empatica E4 wristband on their least dominant hand to record physiological data, which consisted of EDA, BVP and skin temperature. To collect the ground truth data on subjective workload, we used the NASA–TLX questionnaire [15], which participants completed after each puzzle-solving session.

4.3 Procedure

Participants first provided informed consent and completed a demographic form covering age, gender, education, tangram experience, and handedness. After a verbal briefing, they wore a Microsoft HoloLens 2, and an Empatica E4 wristband on their non-dominant hand. The experiment application

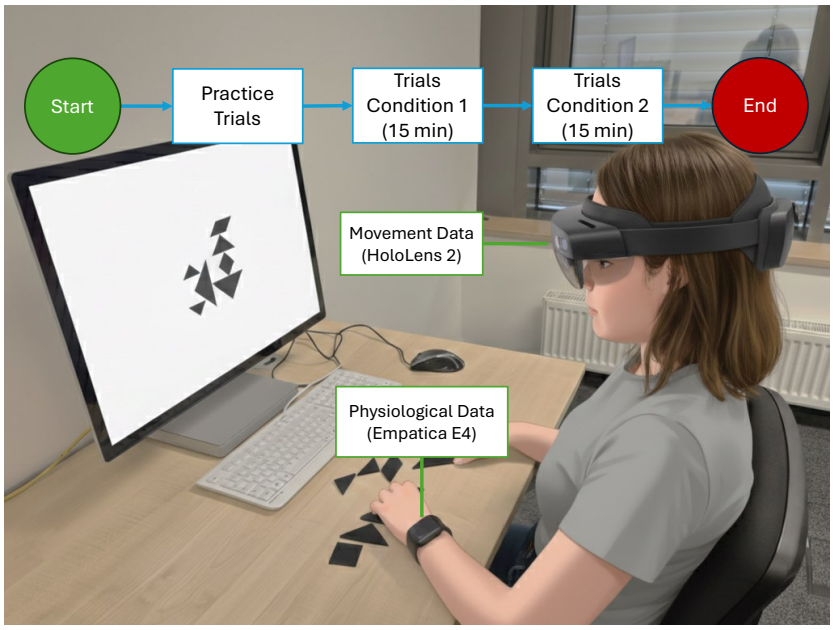


Figure 2: Participants sat in front of a computer and, after a practice session, they solved tangram puzzles of one difficulty (high or low) for 15 minutes, followed by another 15 minutes of puzzles of the other difficulty. During the experiment, they wore a HoloLens 2, which collected movement data, and an Empatica E4, which collected physiological data.

was then started.

The experiment app guided participants through an introduction to tangram, followed by a practice session to familiarize them with puzzle-solving and system interaction. The main task consisted of two 15-minute sessions, one with high and one with low complexity puzzles, where participants solved as many puzzles as possible. To control for ordering effects, we presented the sessions in a counterbalanced order: some participants solved low-complexity puzzles first and others solved high-complexity puzzles first. However the order of the puzzles within each session was constant across participants. If a participant solved all puzzles in a session, the puzzles were repeated in the same order. After each puzzle, participants pressed the space bar to continue; if a puzzle remained unsolved after five minutes, a prompt allowed them to skip it by pressing "S".

During each session, physiological and movement data was collected and participants were asked to look through the AR glasses the whole time, to make sure their eye movements were recorded correctly. After each session, participants completed the NASA–TLX to report their subjective workload and were offered to take a break.

Throughout the experiment, the researcher remained present to monitor progress, provide clarification, and ensure smooth operation. After

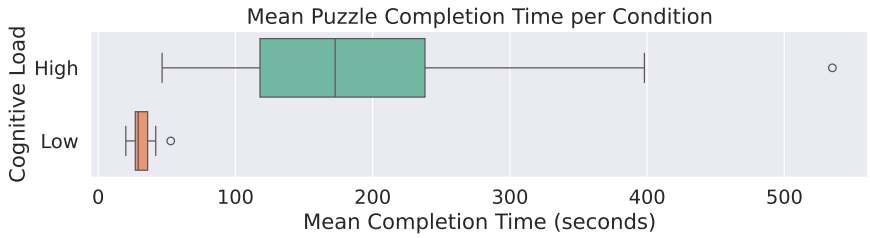


Figure 3: Mean puzzle completion time across high and low complexity conditions.

completing both sessions and questionnaires, the participants were debriefed and thanked for their participation.

4.4 Public Dataset

The dataset consists of a single compressed file (.zip) containing one directory per participant. Each directory contains Comma-Separated Values (CSV) files with time-series data, organized by modality. The files' columns are listed in the first line, beginning with "timestamp", followed by sensor-specific columns. For example, the file "empatica_e4_acc.csv" contains "x", "y", and "z", which represent accelerometer data across three axes or dimensions. In addition to sensor data, the dataset includes two other CSV files. First, "sessions.csv" provides information on each puzzle-solving session, including participant ID, task complexity (0 or 1), and start and end times. The second file, "nasa-tlx_scores.csv", contains all NASA–TLX scores per participant per session, including all sub-scales. The public dataset can be accessed at `https://osf.io/45tjp`.

5 Results

5.1 Experimental Manipulation of Cognitive Load

In this section we present the results of our analyses. We first report task completion times and NASA–TLX scores used to evaluate the experimental manipulation of CL, followed by the results of the CL estimation.

5.1.1 Task Completion Time

The behavioral data we analyzed consisted of puzzle completion times, calculated as the difference in seconds between the start and end of each trial. Based on our design and our preliminary results, we expected puzzles of high complexity to take longer to solve compared to low-complexity puzzles on average.

To test this hypothesis, we calculated average completion times for puzzles in each experimental condition. As shown in Figure 3, the mean completion time for high-complexity puzzles was of $M = 202.53$ seconds ($SD = 127.79$), whereas for low-complexity it was $M = 31.84$ seconds ($SD = 8.45$). To test whether these differences were statistically significant,

we first checked whether the differences between conditions follow a normal distribution. The Shapiro-Wilk test [28] showed that the distribution was not normally distributed ($W = .863$, $p < 0.05$). This was expected, because the underlying phenomenon, task duration, is positively bounded and typically right-skewed. We used the Wilcoxon test, a non-parametric version of the paired T-test[40]. This test showed a statistically significant difference between completion times ($W = 136.0$, $p < .001$), implying that our experimental manipulation of the task led to observable differences in participant behavior. These results indicate that the increased CL associated to high-complexity puzzles substantially affected task efficiency, as measured by task completion times.

5.1.2 NASA-TLX Scores

Subjective outcomes were measured using the NASA–TLX questionnaire, completed by participants after each session. The scale consists of six dimensions, each measured on a 21-point scale ranging from 0–20. We multiplied each value by five, which resulted in values ranging from 0–100. Following recent recommendations that question the mathematical validity of aggregate workload scores [7], we analyzed each sub-scale independently rather than computing a composite score.

To compare perceived workload across conditions, we applied Pratt's variant of the Wilcoxon signed-rank test [24], which is well-suited for ordinal data and handles tied ranks effectively. Participants reported significantly higher scores under the high-complexity condition for *mental demand* ($W = .00$, $p < .001$), *temporal demand* ($W = 51.00$, $p < .05$), *performance* (reverse-scored; $W = .00$, $p < .001$), *effort* ($W = .00$, $p < .001$), *frustration* ($W = 6.00$, $p < .001$). In contrast, *physical demand* did not differ significantly between conditions ($W = 111.00$, $p = .501$). This result was expected, because increasing the complexity of a tangram puzzle does not fundamentally change the physical activity required to solve it. Mean sub-scale scores for each condition are shown in Figure 4.

These findings indicate that the increased CL associated to high-complexity puzzles substantially increased perceived workload, as measured by the NASA–TLX. Together with our findings regarding completion times, these results suggest that our experimental design successfully manipulated CL in the context of our assembly task.

5.2 Cognitive Load Classification

In this study, we trained several machine learning models for binary classification of CL: Logistic Regression (LogReg), Support Vector Machine for Classification (SVC), RF, eXtreme Gradient Boosting (XGB), Light Gradient-Boosting Machine (LightGBM) and Adaptive Boosting (AdaBoost). To improve the ability of the model to generalize to new participants, we used a leave-one-participant-out cross-validation approach, where the data of each participant was used once as the test set, while the rest of the data formed the training set. Standardization (i.e., calculation of the mean and standard deviation) was performed only on the training

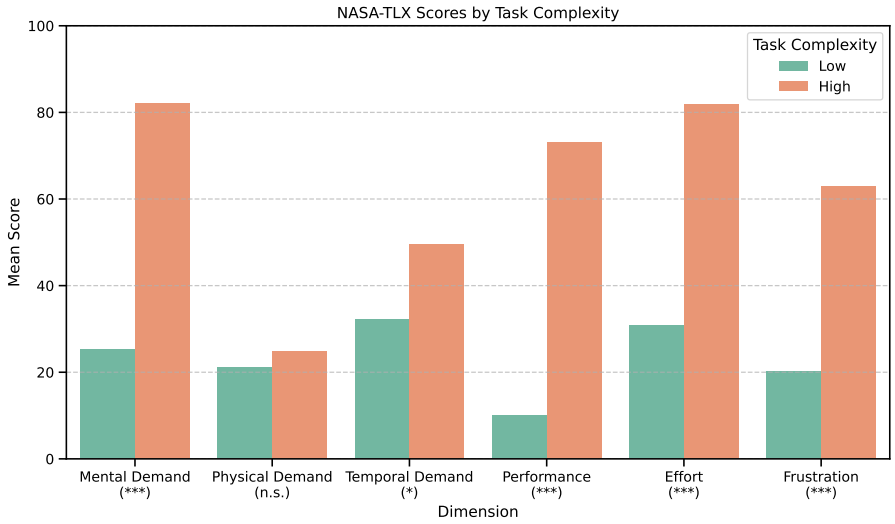


Figure 4: Comparison of experimental conditions across NASA–TLX sub-scales. Higher scores indicate a higher perceived task workload. The scores of the *Performance* sub-scale were reversed, such that higher levels indicate reduced feelings of success when performing the task, and thus a higher workload. Significant differences between experimental conditions are indicated below each NASA–TLX sub-scale: (***) $p < .001$, (*) $p < .05$, (n.s.) $p >= .05$.

data within each fold, and these parameters were then applied to the held-out test participant.

To analyze the contribution of each sensor modality, we trained and evaluated our models on three distinct datasets: movement measures (recorded by the HoloLens 2), physiological measures (recorded by the Empatica E4), and a combined (multimodal) dataset integrating both. Model performance was evaluated using accuracy (the percentage of correctly classified samples) and macro F1 scores. We prioritized the macro F1 metric to ensure a balanced assessment of precision and recall across both the high and low complexity classes, preventing potential over-reliance on a majority class.

Our best-performing model achieved 88.6% F1 score and accuracy when trained on movement data alone (i.e., eye, head, and hand features) using RF. In contrast, models using only physiological data (EDA, BVP and skin temperature) reached a maximum of 59% F1 score and accuracy, indicating a weaker signal for CL classification in our setup. Combining physiological and movement data yielded slightly lower performance than using movement data alone ($F1 = .831$), a pattern also observed by Hou et al. [17]. This may indicate that combining modalities does not always lead to better performance, potentially due to noise in physiological features or mismatched signal patterns across participants. Table 2 shows a detail comparison of all classifier performances across modalities.

Table 2: F1 score and accuracy performance per model and modality using leave-one-participant-out cross-validation.

	Movement measures		Physiological measures		Movement + Physiological	
Model	F1	Acc.	F1	Acc.	F1	Acc.
AdaBoost	0.866	0.867	0.572	0.574	0.809	0.809
LGBM	0.882	0.882	0.514	0.516	0.826	0.826
LogReg	0.833	0.834	**0.590**	**0.591**	0.812	0.812
RF	**0.886**	**0.886**	0.514	0.514	0.826	0.826
SVC	0.856	0.856	0.574	0.575	0.773	0.774
XGB	0.882	0.882	0.469	0.572	**0.831**	**0.832**

Notes: Physiological measures consist of EDA, BVP and skin temperature, whereas movement measures consist of eye, head and hand movements.

Table 3: Feature importance of movement data by modality.

Feature Modality	*N*	Total Importance	Top Feature
Eye movements	90	54.1%	fixation_distance_mean
Hand movements	176	23.5%	hand_max_position_change_iqr
Head movements	45	22.4%	head_up_angular_distance_mean

Notes: *N* denotes the number of features within each modality. Total importance represents the mean cumulative Gini importance across all participants in the leave-one-participant-out cross-validation, normalized to 100%.

To understand what drove the high performance of movement data, we analyzed the relative importance of features in the best performing model (Table 3). The model was trained using RF with leave-one-participant-out cross-validation, and feature importance was calculated as the mean cumulative Gini importance across all participants, normalized to 100%. We found that eye movements, comprising features related to fixations, saccades and blinks, accounted for 54.1% of the total importance, followed by hand movements (23.5%) and head movements (22.4%). This suggests that eye movements were the most important feature for CL classification in our setup.

6 Discussion

In this study, we explored the feasibility of machine learning-based CL classification using AR device sensors in assembly contexts. Our findings showed that the best CL classification performance ($F1 = .886$) was achieved when using movement data alone, consisting of head, hand, and eye movements. This performance surpassed both the combined dataset and the physiological data on their own (consisting of EDA, skin temperature and BVP), which led to the worst classification performance. This provides the first strong empirical evidence that AR headsets can function as standalone tools for CL estimation in assembly tasks, minimizing intrusiveness and resource consumption.

Our best F1 performance achieved with a lightweight RF model, establishes a strong benchmark for resource-efficient CL estimation on AR devices. While Hou et al. [17] reported a higher F1 score of 96%, their result relied on a resource-intensive Transformer architecture and was likely influenced by data leakage, as noted in Section 2 (Related Work). Moreover, a direct comparison is limited due to methodological differences: our approach utilized a binary classification scheme (instead of three classes) and was validated using rigorous leave-one-participant-out cross-validation, ensuring better generalizability to unseen participants. Furthermore, our task is closer to a manual assembly task than the button-pressing task used in Hou et al. [17].

In analyzing the features that contributed most to the highest-performing model, we found that eye-movement features accounted for more than half of the total importance. This result aligns with previous research describing eye movements as the modality most sensitive to intrinsic CL, when compared to physiological measures, such as heart-rate variability [3]. However, we identified a gap in the literature regarding the relative importance of eye movements compared to head and hand kinematics in CL classification. We hypothesize that prominence of eye movements in this study stems from the task's heavy reliance on visual attention. This is consistent with CLT [33], as the visual route serves as the primary channel for information acquisition during manual assembly. Consequently, eye movements likely represent the most sensitive modality for estimating CL in similar visuospatial tasks. Nevertheless, head and hand features remained significant predictors: head movements likely supported visual gaze shifts, while hand movements directly reflected the physical execution and motor planning required by the assembly process.

Our observation that model performance decreased when adding physiological data to the movement data is significant. This pattern was also observed by Hou et al. [17]. Physiological data is considered valid predictor of cognitive load [3]. But, since physiological measures rely on biological responses, it is possible that these measures may be more effective when used to detect tasks which increase biological responses as demands increase. For example, in Hou et al. [17], the "Physical Demand" component of NASA–TLX increased in correlation with noise levels. However, in our task, "Physical Demand" showed no significant change between conditions, which could affect the ability of the models to detect CL based on biomarkers. In this context, adding physiological data could introduce noise to the model instead of contributing with useful signal. Moreover, the limitations of physiological data have been emphasized in the literature, for example, Vanneste et al. [35] reported a fairly low explained variance of 22.8% on their model for CL based solely on physiological measures, which could explain our low performance in our physiological-only model. On the other hand, movement data directly reflect task engagement and motor execution, offering a more accurate prediction, as observed in our results.

7 Limitations and Future Work

While models based solely on movement data achieved the best performance in our study, this reliance introduces a trade-off regarding the model's ecological validity. Movement patterns are mostly task-dependent; in our study, participants were limited to a seated, tabletop tangram task designed to mimic assembly scenarios with similar characteristics. In contrast, real-world manual assembly tasks can vary widely and often encompass a broader range of motion, such as walking or operating heavy machinery. Such diverse physical activities may introduce kinematic variance that differs significantly from our controlled setup, potentially limiting the direct transferability of models trained on these specific motor patterns.

Our results suggest that movement data is a highly effective predictor of CL in our experimental scenario. However, its robustness in unstructured or physically dynamic environments remains to be established. While physiological data, such as HR and EDA, could theoretically offer better generalization across diverse assembly scenarios, these signals are also susceptible to the confounding effects of physical exertion. Future research should, therefore, focus on methods to decouple the physiological and kinematic signatures of physical exertion from those of CL to determine the extent to which these models can generalize to unconstrained, real-world industrial settings.

A second limitation of our study is the operationalization of cognitive load. Our two complexity conditions (high and low) were highly distinct, as evidenced by participants' NASA–TLX scores and mean completion times. However, such a stark contrast may have oversimplified the classification task, rendering it less representative of real-world CL variations. Future research should consider introducing a third, intermediate level of cognitive load, potentially by further manipulating visual complexity (e.g., increasing the number of visible contours, as proposed by Vanneste et al. [35]), or narrowing the gap between the existing levels. This would better approximate the subtle differences in CL encountered in practical assembly scenarios, particularly those supported by AR or traditional instruction methods.

Beyond classification performance, our window-based approach highlights the potential for developing adaptive AR systems. By partitioning the data into 30-second segments, we demonstrate that movement-based features can provide timely detection of CL fluctuations. This capability is a crucial prerequisite for systems that dynamically adjust instructional support based on the user's current cognitive state. Future research should explore the integration of these models into adaptive pipelines, optimizing window durations to balance detection latency with classification accuracy for seamless instructional interventions.

8 Conclusion

In summary, this study provides the first empirical evidence that movement data captured solely from AR headsets can serve as an effective proxy for CL during manual assembly tasks, achieving an F1 score of 0.886 in

binary classification. While these findings are currently bounded by the seated nature of the task and the binary complexity levels, they highlight a clear path toward less intrusive CL monitoring. Ultimately, this approach enhances the ecological validity of AR-supported instruction by simplifying experimental setups and providing a foundation for adaptive assistance in industrial settings.

Acknowledgments

This work was supported by MEDIUS (German Federal Ministry of Education and Research, code 02P20A054). In addition, this work was partially supported by Tateisi Science and Technology Foundation Research Grant (A). Gemini was used to anonymize Figure 2 and provided suggestions for textual refinements. All AI-generated recommendations were critically reviewed and adapted by the co-authors. We thank the participants for their time and effort.

Author Contribution

Javier Melo: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Project administration.
Leyla Akinci: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing.
Ko Watanabe: Writing - review & editing, Supervision.
Nicolas Großmann: Writing - review & editing, Supervision, Funding acquisition.
Shoya Ishimaru: Writing - review & editing, Supervision, Funding acquisition.
Andreas Dengel: Resources, Supervision, Funding acquisition.

References

[1] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(3):1–20, 2017.

[2] Lakmal Abeysekera and Phillip Dawson. Motivation and cognitive load in the flipped classroom: definition, rationale and a call for research. Higher education research & development, 34(1):1–14, 2015.

[3] Paul Ayres, Joy Yeonjoo Lee, Fred Paas, and Jeroen JG Van Merrienboer. The validity of physiological measures to identify differences in intrinsic cognitive load. Frontiers in psychology, 12:702538, 2021.

[4] Samantha D. Aziz and Oleg V. Komogortsev. An Assessment of the Eye Tracking Signal Quality Captured in the HoloLens 2. In 2022

Symposium on Eye Tracking Research and Applications, pages 1–6, June 2022. doi: 10.1145/3517031.3529626. URL http://arxiv.org/abs/2111.07209. arXiv:2111.07209 [cs].

[5] Francesco N Biondi, Angela Cacanindin, Caitlyn Douglas, and Joel Cort. Overloaded and at work: investigating the effect of cognitive workload on assembly task performance. Human factors, 63(5):813–820, 2021.

[6] Jonas Blattgerste, Benjamin Strenge, Patrick Renner, Thies Pfeiffer, and Kai Essig. Comparing conventional and augmented reality instructions for manual assembly tasks. In Proceedings of the 10th international conference on pervasive technologies related to assistive environments, pages 75–82, 2017.

[7] Matthew L Bolton, Elliot Biltekoff, and Laura Humphrey. The mathematical meaninglessness of the nasa task load index: A level of measurement analysis. IEEE Transactions on Human-Machine Systems, 53(3):590–599, 2023.

[8] Georg Buscher, Andreas Dengel, and Ludger Van Elst. Eye movements as implicit relevance feedback. In CHI '08 Extended Abstracts on Human Factors in Computing Systems, pages 2991–2996, Florence Italy, April 2008. ACM. ISBN 978-1-60558-012-8. doi: 10.1145/1358628.1358796. URL https://dl.acm.org/doi/10.1145/1358628.1358796.

[9] Lea M Daling and Sabine J Schlittmeier. Effects of augmented reality-, virtual reality-, and mixed reality–based training on objective performance measures and subjective evaluations in manual assembly tasks: a scoping review. Human factors, 66(2):589–626, 2024.

[10] Juan C Dibene and Enrique Dunn. Hololens 2 sensor streaming. arXiv preprint arXiv:2211.02648, 2022.

[11] Marshall G. Doane. Interaction of Eyelids and Tears in Corneal Wetting and the Dynamics of the Normal Human Eyeblink. American Journal of Ophthalmology, 89(4):507–516, April 1980. ISSN 0002-9394. doi: 10.1016/0002-9394(80)90058-6. URL https://linkinghub.elsevier.com/retrieve/pii/0002939480900586. Publisher: Elsevier BV.

[12] Halimoh Dorloh, Kai-Way Li, and Samsiya Khaday. Presenting job instructions using an augmented reality device, a printed manual, and a video display for assembly and disassembly tasks: what are the differences? Applied Sciences, 13(4):2186, 2023.

[13] Paul Evans, Maarten Vansteenkiste, Philip Parker, Andrew Kingsford-Smith, and Sijing Zhou. Cognitive load theory and its relationships with motivation: A self-determination theory perspective. Educational Psychology Review, 36(1):7, 2024.

[14] Johannes Funk and Ludger Schmidt. Evaluation of an augmented reality instruction for a complex assembly task: comparison of a smartphone-based augmented reality instruction with a conventional paper instruction for the teach-in phase in manual assembly. i-com, 20(1):63–72, 2021.

[15] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Advances in psychology, volume 52, pages 139–183. Elsevier, 1988.

[16] Lei Hou, Xiangyu Wang, and Martijn Truijens. Using augmented reality to facilitate piping assembly: an experiment-based evaluation. Journal of Computing in Civil Engineering, 29(1):05014007, 2015.

[17] Yukang Hou, Qingsheng Xie, Ning Zhang, and Jian Lv. Cognitive load classification of mixed reality human computer interaction tasks based on multimodal sensor signals. Scientific Reports, 15(1):13732, 2025.

[18] Jiali Liang, Krista Wilkinson, and Robert L Sainburg. Is hand selection modulated by cognitive–perceptual load? Neuroscience, 369: 363–373, 2018.

[19] Deepti Majumdar, Kiran Mondal, and Tammanna R Sahrawat. Eye movement metrics as indicator of cognitive loading: A systematic review. Research Aspects in Biological Science, 2:1–17, 2022.

[20] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods, 53(4):1689–1696, August 2021. ISSN 1554-3528. doi: 10.3758/s13428-020-01516-y. URL `https://link.springer.com/10.3758/s13428-020-01516-y`.

[21] Javier Melo, Leigh Fernandez, and Shoya Ishimaru. Automatic classification of difficulty of texts from eye gaze and physiological measures of l2 english speakers. IEEE Access, 2025.

[22] Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. Psychopy2: Experiments in behavior made easy. Behavior research methods, 51:195–203, 2019.

[23] Polypad. Tangram builder. `https://polypad.amplify.com/tangram`, n.d. Accessed: 2025-05-27.

[24] John W. Pratt. Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures. Journal of the American Statistical Association, 54(287):655–667, September 1959. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1959.10501526. URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.1959.10501526`.

[25] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the symposium on Eye tracking research & applications - ETRA '00, pages 71–78, Palm Beach Gardens, Florida, United States, 2000. ACM Press. ISBN 978-1-58113-280-9. doi: 10.1145/355017.355028. URL http://portal.acm.org/citation.cfm?doid=355017.355028.

[26] Seyed Ehsan Samaei, Shahram Vosoughi, Ebrahim Taban, Majid Bagheri Hossein Abadi, Ghasem Zia, and Mohammad Hossein Beheshti. The effect of mental workload on occupational accidents among nurses in hospitals of kerman, iran. International Journal of Hospital Research, 6(4):63–75, 2017.

[27] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. Discriminating stress from cognitive load using a wearable eda device. IEEE Transactions on information technology in biomedicine, 14(2):410–417, 2009.

[28] S. S. Shapiro and M. B. Wilk. An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4):591, December 1965. ISSN 00063444. doi: 10.2307/2333709. URL https://www.jstor.org/stable/2333709?origin=crossref.

[29] Soroosh Solhjoo, Mark C Haigney, Elexis McBee, Jeroen JG van Merrienboer, Lambert Schuwirth, Anthony R Artino Jr, Alexis Battista, Temple A Ratcliffe, Howard D Lee, and Steven J Durning. Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. Scientific reports, 9(1): 14668, 2019.

[30] Yuko Suzuki, Fridolin Wild, and Eileen Scanlon. Measuring cognitive load in augmented reality with physiological methods: A systematic review. Journal of Computer Assisted Learning, 40(2):375–393, 2024.

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

[32] John Sweller. Cognitive load theory, learning difficulty, and instructional design. Learning and instruction, 4(4):295–312, 1994.

[33] John Sweller. Cognitive Load Theory: Recent Theoretical Advances. In Jan L. Plass, Roxana Moreno, and Roland Brünken, editors, Cognitive Load Theory, pages 29–47. Cambridge University Press, 1 edition, April 2010. ISBN 978-0-521-67758-5 978-0-521-86023-9 978-0-511-84474-4. doi: 10.1017/CBO9780511844744.004. URL https://www.cambridge.org/core/product/identifier/CBO9780511844744A012/type/book_part.

[34] Saki Tanaka, Airi Tsuji, and Kaori Fujinami. Eye-tracking for estimation of concentrating on reading texts. International Journal of Activity and Behavior Computing, 2024(1):1–21, 2024. doi: 10.60401/ijabc.10.

[35] Pieter Vanneste, Annelies Raes, Jessica Morton, Klaas Bombeke, Bram B Van Acker, Charlotte Larmuseau, Fien Depaepe, and Wim Van den Noortgate. Towards measuring cognitive load through multimodal physiological data. Cognition, Technology & Work, 23:567–585, 2021.

[36] Xiangyu Wang, Soh K Ong, and Andrew YC Nee. A comprehensive survey of augmented reality assembly research. Advances in Manufacturing, 4:1–22, 2016.

[37] Ko Watanabe, Tanuja Sathyanarayana, Andreas Dengel, and Shoya Ishimaru. Engauge: Engagement gauge of meeting participants estimated by facial expression and deep neural network. IEEE Access, 11:52886–52898, 2023. doi: 10.1109/ACCESS.2023.3279428.

[38] Ko Watanabe, Andreas Dengel, and Shoya Ishimaru. Metacognition-engauge: Real-time augmentation of self-and-group engagement levels understanding by gauge interface in online meetings. In Proceedings of the Augmented Humans International Conference 2024, AHs '24, page 301–303, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400709807. doi: 10.1145/3652920.3653054. URL https://doi.org/10.1145/3652920.3653054.

[39] Stefan Wiedenmaier, Olaf Oehme, Ludger Schmidt, and Holger Luczak. Augmented reality (ar) for assembly processes design and experimental evaluation. International journal of Human-Computer interaction, 16(3):497–514, 2003.

[40] Frank Wilcoxon. Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6):80, December 1945. ISSN 00994987. doi: 10.2307/3001968. URL https://www.jstor.org/stable/10.2307/3001968?origin=crossref.