Cognitive Load Inference Using Physiological Markers in Virtual Reality

Jishang Wei* Erika Siegel† Prahalathan Sundaramoorthy‡ Antônio Gomes§ Shibo Zhang¶
Mithra Vankipuram‖ Kevin Smathers** Sarthak Ghosh†† Hiroshi Horii‡‡ Jeremy Bailenson
Rafael 'Tico' Ballagas

HP Labs
Stanford University

ABSTRACT

Virtual reality (VR) has become an increasingly popular way for learning and training. The assessment of the amount of mental effort, or cognitive load required to perform a task, is essential to create adaptive VR experiences. In this work, we conducted a large-scale study (N=738) to collect behavioral and physiological measures under different cognitive load conditions in a VR environment, and developed a novel machine learning solution to predict cognitive load in real time. Our model predicts cognitive load as a continuous value in the range from 0 to 1, where 0 and 1 correspond to the lowest and highest reported cognitive loads across all participants. On top of the point estimation of cognitive load, our model quantifies prediction uncertainty using a prediction interval. We propose a novel dual-branch attention model to accurately predict the cognitive load. We achieve a MAE (Mean Absolute Error) of 0.11. The result indicates that, with a combination of behavioral and physiological indicators, we can reliably predict cognitive load in real-time, without calibration. To support further research, we are releasing a test dataset comprising data from 100 participants for use by researchers and developers interested in machine learning, virtual reality, learning & memory, cognition, or psychophysiology. This dataset includes recordings from multiple sensors (including pupillometry, eye-tracking, and pulse plethysmography), self-reported cognitive effort, behavioral task performance, and demographic information on the sample.

Index Terms: Virtual Reality, Cognitive Load, Physiological Signals, Machine Learning

1 INTRODUCTION

Virtual reality (VR) is revolutionizing learning and training by offering individuals the ability to safely engage in novel experiences from their home or office. Cognitive load, which refers to the mental effort required to process information, often interferes with the fundamental goals people have when using VR, such as immersive learning or skill development. High cognitive load can overwhelm users, impairing their ability to retain information and complete tasks efficiently, especially in educational settings [33]. Research on immersive virtual reality learning demonstrates how factors such as presence, user agency, cognitive and emotional influences, and haptic feedback shape learning outcomes in these environments [33, 43]. In our work, we aim to introduce a new generation of adaptive VR training tools that personalize the experience in real-time, responding to the individual's mental state. To bring this vision to life, we are developing an AI-driven "inference engine" that uses physiological sensors embedded in the head-mounted display to monitor and interpret cognitive load in VR.

Cognitive load, has been studied by researchers interested in learning and performance for over a century [11, 37, 56, 59, 63]. Every person has their own information processing capacity (also called working memory capacity or short-term memory), and it is fixed (unchanging) [4], limited (small capacity) [14, 42], and varies from person to person [17, 41]. Overwhelming our cognitive capacity hampers effective processing. If the load is too much, comprehension stalls; if too little, it may not be worth the efforts.

Physiological measures such as pupillometry, eye tracking and cardiac activity have been proven to be useful indicators of emotion, sentiment, and cognitive load [28, 29, 50, 53]. In our study, we aim to capture this information from reliable sensors integrated in a head-mounted display while participants perform various tasks requiring different amounts of cognitive load in a VR setting. Our goal is to develop a solution that could reliably predict cognitive load in the general population through physiological signals. We designed a study with a series of tasks to stimulate different levels of cognitive load from participants and passively recorded their physiology, tracked their performance, and collected self-report data. Our approach differed from past research in a number of ways, and included novel innovations from other "cognitive workload" solutions. First, our sample is larger (N=738) and more diverse (across four continents) than previous studies. To minimize sampling bias, we intentionally chose a diverse population with variance in age, educational level, and race/ethnicity rather than a homogeneous one to account for individual differences [44, 47]. Second, our feature selection process focused on robustness versus parsimony. It was more important that we could reliably predict cognitive load under a variety of user and task conditions. Third, we wanted a real-time, calibration-free solution, which required important innovations in signal processing, labeling, and feature engineering. In general, commercial systems that predict mental workload do so post-hoc, during some type of after-action-review (e.g., a user's cognitive load is estimated by subtracting responses during a task from some type of baseline or calibration task).

We design a machine learning model and train it with paired input physiological signals and cognitive load labels. The model fuses information across the signals to predict a probabilistic distribution of cognitive load. The mostly likely cognitive load value with an interval is derived from the distribution to represent the final prediction result. Cognitive load is a complex construct which is related to both individual cognitive capacity and the task being performed. Particularly, different people may experience different cognitive load while performing the same task and the experi-

*e-mail: jishang.wei@hp.com
†e-mail: erika.siegel@hp.com
‡e-mail: prahalathan.sundaramoorthy@hp.com
§e-mail: antonio.gomes@hp.com
¶e-mail: shibo.zhang@hp.com
‖e-mail: mithra.vankipuram@hp.com
**e-mail: kevin.smathers@hp.com
††e-mail: sarthak.ghosh@hp.com
‡‡e-mail: hiroshi.horii@hp.com
e-mail: bailenson@stanford.edu
e-mail: tico.ballagas@hp.com

enced cognitive load changes across tasks of various difficulty levels. Given this, we use both subjective cognitive load ratings and task difficulties as labels to train the model. Our solution is unique on a few aspects. First, our model predicts real-valued cognitive load while the majority of other cognitive load prediction solutions predict only the coarse levels of cognitive load. Second, our model is designed and trained with dual targets, subjective cognitive load and objective task difficulty, which implicitly captures the mixture distribution of cognitive load in different context. Third, rather than outputting a single value, we estimate a probabilistic distribution of cognitive loads to quantify prediction uncertainty due to individual difference in cognitive load manifestation.

Our cognitive load "inference engine" has been integrated into commercial VR headsets and is widely utilized by researchers and developers across academia and industry in a variety of VR applications. With this paper, we release a subset of our dataset (N=100)[1] with the larger scientific community for validation of our models and to seed new research ideas. We hope researchers in the field of virtual reality, psychophysiology, cognitive, and learning will find these data useful and fruitful.

2 Related Work

2.1 Cognitive Load

Research shows that cognitive load is an important predictor of learning, memory, performance, stress, and burnout [3, 24, 35, 42, 55]. An examination of the theoretical underpinnings of cognitive load, particularly cognitive load theory [56], can provide insight into why. Cognitive load theory suggests that successful completion of any task (large or small) relies on the complex interplay between sensory inputs, long-term memory (acting as a repository of previously acquired knowledge and skills), and working memory. Working memory acts as an intermediate state between sensory and long-term memory, attaching meaning to the sensory information by integrating newly learned information into longer-term memory. Both sensory and long-term memories have flexible capacities and are capable of processing large volumes of information. Working-memory, on the other hand, is comparatively limited [4, 17]. Attention manages the function of working memory by guiding it to relevant sensory information and stored knowledge, thereby directing the learning process and increasing (or decreasing) the efficiency of working memory [41]. An individual's cognitive load in each moment is an amalgam of these attentional, sensory, and memory processes. When we measure cognitive load, we are estimating the amount of mental resources being utilized to complete the task at hand.

2.2 Indicators of Cognitive Load

The quest to understand human cognition has driven researchers to seek reliable and objective indicators of cognitive load, bypassing the need to inquire about participants' experiences.

Measures of the peripheral nervous system and eye-tracking behavior serve as more reliable and generalizable indicators of mental effort [20]. Cardiovascular measures such as blood pressure [46], heart rate [18, 19, 21, 26], and high-frequency heart rate variability [15, 18] have been shown to reliably estimate changes in cognitive load levels. Similarly, electrodermal activity (EDA), also referred to as skin conductance [23,51], and pupillometry, the measure of pupil dilation, have been proven to be reliable indicators [42]. Multiple studies have established a strong correlation between task demands and pupil dilation [9, 31]. Additionally, electroencephalography (EEG) is a non-invasive method that captures the brain's electrical activity. By examining specific EEG patterns and frequencies, researchers can gauge the brain's responses to different cognitive demands [12,62]. Certain brainwave frequencies might intensify or diminish based on the cognitive load, making EEG a valuable tool for understanding the nuances of human cognition under varying challenges.

Some researchers have attempted to estimate cognitive load unobtrusively by recording and categorizing participants' facial expressions during the task [1, 9, 52]. Unfortunately, facial expressions have been unreliable indicators for both practical (dynamic facial expressions can be difficult to infer [13]) and theoretical (the categories themselves are up for debate [7]) reasons. Researchers have explored acoustic features of the voice [10] to predict cognitive load. However, findings do not reliably generalize across cognitive load tasks [54]. Moreover, individuals don't always vocalize during tasks.

2.3 Cognitive Load Inference with Machine Learning

Physiological indicators sensitive to cognitive load enable real-time cognitive load estimation. Researchers have developed various machine learning algorithms for this purpose, leveraging features from different physiological signals. Commonly used algorithms include k-nearest neighbor (kNN), naïve Bayes (NB), logistic regression, linear discriminant analysis (LDA), support vector machines (SVM), ensemble methods (e.g., random forest, XGBoost), and neural networks. These machine learning models are trained to predict users' cognitive load based on physiological features from one or multiple signal modalities. For example, Nourbakhsh et al. [40] used SVM and NB to train a cognitive load prediction model based on skin conductance and blink features. Haapalainen et al. [21] trained NB to classify three levels of cognitive load using electrocardiogram (ECG) and eye movement features. Jimenez-Molina et al. [27] employed logistic regression, SVM, and neural network models based on features from EDA, ECG, pulse plethysmography (PPG), EEG, temperature, and pupil dilation signals.

Cognitive load and psycho-physiological research indicate that different signal modalities provide complementary and overlapping information. For example, increased cognitive load may cause accelerated heart rate with reduced variability [2, 21], pupil dilation [42], more saccades [9], or increased blink activity [45]. Multi-modal fusion enhances cognitive load prediction by integrating these correlated physiological responses, offering a clearer picture of autonomic nervous system activity than any single signal. Additionally, fusion improves robustness by reducing noise effects and capturing responses that unfold at different timescales, such as pupil dilation and pulse transit time.

Common multi-modal fusion strategies in machine learning include feature-level fusion (early fusion), decision-level fusion (late fusion), and hybrid fusion. Studies such as [2, 18, 27] have used feature-level fusion to combine features from different signals, while Islam et al. [25] merged features to predict drivers' mental workloads. Zhang et al. [64] explored decision fusion for more robust predictions, and also experimented with hybrid fusion. Recently, neural networks have shown promise for cognitive load prediction. Sarkar et al. [49] used a multitask deep neural network with ECG signals, and Saha et al. [48] applied LSTM and MLP with EEG signals. These studies indicate that deep learning models outperform traditional approaches like kNN, SVM, and LDA.

3 Cognitive Load Study and Data Collection

3.1 Consent for Human Subjects Research

We implemented a comprehensive internal review process to ensure ethical research practices and participant protection, given that this research fell outside the scope requiring formal Institutional Review Board (IRB) oversight at our institution. Our protocol underwent thorough evaluation by HP Legal and HP Privacy to verify compliance with data protection regulations and research ethics standards. All participants provided informed consent prior to study

[1] data access link: https://developers.hp.com/omnicept/read-me-cognitive-load-open-dataset

Table 1: Demographic Characteristics of the Sample

Demographic Category	Percentage of Sample
Education	
Elementary School	4%
High School	25%
College	61%
Graduate Degree	10%
Age Range	
18-24	17%
25-34	42%
35-49	34%
50-64	6%
Gender	
Female	48%
Male	52%
Race/Ethnicity	
Caucasian	81%
African	9%
Asian	1%
Multiracial	9%

enrollment, with clear documentation of how their data would be collected, used, and protected. The consent process detailed participants' rights, including the ability to withdraw from the study at any time and have their data removed. Data handling procedures incorporated privacy-preserving measures, including data minimization, secure storage protocols, and restricted access controls. All personally identifiable information was encrypted and stored separately from research data using industry-standard security measures.

3.2 Participants

738 participants were recruited from various communities for this experiment. The age range of the participants was between 19 and 61. Participants did not report ophthalmological conditions (other than corrected vision). Participants received a payment in local currency as remuneration for participating in the study. All participants received detailed information about the nature of the study, their role, and how their data would be managed, stored, used to develop new products, and published as a part of this open dataset. Each participant gave their informed consent.

3.3 Design and Procedure

Participants completed a series of tasks presented in a randomized order, designed to require different levels of mental effort, or cognitive load (CL), to complete. Three CL levels were manipulated (*low*, *medium*, and *high*), and each level was repeated three times (3 low trials, 3 medium trials, 3 high trials), in a random order, for a total of nine trials (Fig. 1). At the end of each trial, participants rated how mentally demanding they found the task (Fig. 1). We drew inspiration for the cognitive load stimuli from the study outlined in [8]. The tasks, spanning multiple modalities, were designed to induce cognitive load across a spectrum from the easiest to the most challenging, reflecting real-world scenarios.

In a low CL trial, participants completed a visual vigilance task in which five balls appear in the scene. The ball to be tracked is briefly highlighted in a different color. The ball then move in a random, diverging pattern around the screen, and eventually settle in one of five spots on the screen (labelled A through E, see Fig. 1 A). The users need to indicate where the target ball landed. In the medium CL trial, participants completed the same visual vigilance task while also performing an arithmetic task. Numbers would progressively be presented to the screen while the balls were in motion

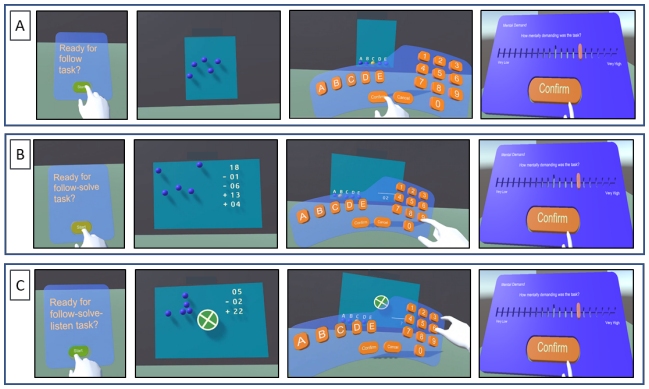

Figure 1: Example stimuli from the cognitive load task. In each trial, participants perform tasks designed to induce varying levels of mental effort: low, medium, and high. Box A illustrates a low-effort trial, where participants track a single ball moving randomly on the screen and report which one they tracked once it stops. Box B shows a medium-effort trial, where participants track a ball and solve mental math problems appearing randomly, then report both the tracked ball and their math answer. Box C depicts a high-effort trial, where participants track a ball, solve math problems, and monitor a spinning wheel, reporting its direction when it stops. After each trial, participants rate the mental demand of the task on a continuous scale from very low to very high. We use a unidimensional scale subjective report due to its simplicity, which makes it easier for participants to provide self-reports while requiring significantly less time. This approach minimizes interruptions between cognitive load trials.

(see Fig. 1 B). The numbers would disappear when prompted for a response, which forced the users to perform the arithmetic task in parallel with the visual vigilance task. In the high CL trial, a third audio vigilance task was added (see Fig. 1 C). In this condition, the subject is listening for an audio beep. An additional visual element of a spinning wheel is layered into the scene, and the direction of spinning randomly changes throughout the task. When the audio prompt occurs, the subject should indicate using the corresponding controller trigger which direction the green wheel is spinning.

The CL tasks are designed so that each level of task difficulty is objectively more challenging, with the more difficult conditions fully encompassing the easier ones. By layering these tasks and requiring participants to perform them simultaneously, attention is divided across tasks, thereby increasing the workload. In the low difficulty condition, participants only engaged in the visual vigilance task. In the medium difficulty condition, they performed two tasks simultaneously: visual vigilance and arithmetic. In the high difficulty condition, participants managed three tasks simultaneously: visual vigilance, arithmetic, and audio vigilance. This layered multi-tasking approach ensures an increase in task difficulty. Performing a single task (low difficulty) is objectively easier than performing the same task concurrently with a second task (medium difficulty). Similarly, managing two tasks simultaneously (medium difficulty) is objectively easier than managing the same two tasks alongside an additional third task (high difficulty).

We have also designed tasks with neutrality in mind in order to best mitigate demographic and individual factors (e.g., age, gender, ethnicity, education) that can influence cognitive load. These tasks were selected to minimize biases related to education or age, focusing on universal cognitive processes such as memory and pattern recognition.

Participants began the task with several practice trials to introduce them to the procedure. The practice trials started with a low

CL trial. If participants failed to complete the task successfully on the first practice trial, they were given two more tries to complete the task. If successful, participants moved on to medium CL, followed by high CL practice trials. If a participant failed to successfully complete a practice trial after three tries, they skipped to the next trial. This feature was added to ensure that participants had relatively equal exposure to the stimuli and were in VR for roughly the same amount of time (minimizing effects from fatigue and/or VR sickness).

In the testing phase, participants were shown a total of nine tasks (three low, three medium and three high) in a randomized order to minimize carry over effects from the last trial. Because participants were, at this point, familiar with the stimuli and virtual reality, we expect that cognitive load effects from the testing phase are the result of our manipulations.

3.4 Setup and Materials

Participants completed a VR experience that was designed to stimulate different levels of cognitive load while we passively recorded data from multiple sensors. The experience was developed by the authors in Unity3D and included a series of tasks similar to the desktop-based stimulus media proposed by Bartels and Marshall [8], but modified for VR.

3.5 Hardware VR Headsets

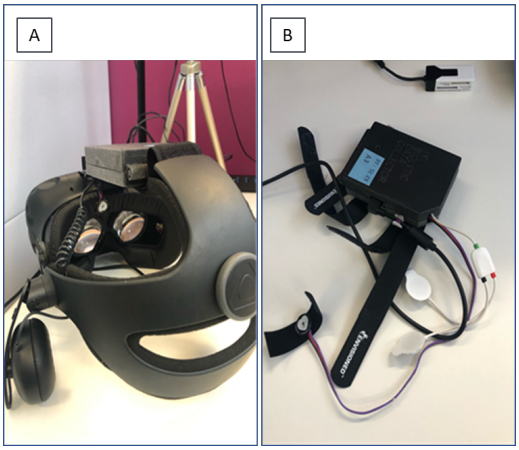

Figure 2: Sensor and Head-Mounted Display (HMD) set up for data collection at the incubation stage. Box A features a modified Vive Pro-Eye HMD with eye-tracking and pupillometry, along with a PPG sensor attached between the lenses. Box B shows the PPG sensor used to record data from the finger.

In the first stage, we used the HTC Vive Pro-eye head mounted display with dual OLED 3.5" diagonal screen, 1440 x 1600 (pixel per eye) resolution, 90Hz refresh rate, 110° field-of-view. The Vive ProEye includes Tobii eye tracking (120Hz gaze output) and pupillometry capabilities [2]. To measure cardiac activity, we used a BITalino (r)evolution wired pulse plethysmography (PPG) sensor (bitalino.com). PPG was collected from the forehead, using a PPG sensor affixed directly to the mask (Fig. 2, Box A) and from the finger (Fig. 2, Box B).

In the second stage, we are able to use HP Reverb G2 Omnicept Edition VR headset, which includes a state-of-the-art sensor system that measures gaze, pupil size and pulse, and seamlessly transfers data to the HP Omnicept VR Software platform.

Participants completed the experience described in 3.4 using the above hardware. During the task, we collected performance from each task and, at the end of each trial, subjective ratings of cognitive load (Fig. 1). Using a modified version of the NASA Task Load Index (TLX), participants rated how "mentally demanding" they found the task on a continuous scale from very low to very high [22].

[2]https://www.vive.com/us/product/vive-pro-eye/specs/

4 Data Processing

4.1 Pupilometry and Eye Tracking Data

We measured dilation of the pupil and tracked participants' eyes using a Tobii Eye Tracking system integrated in the HTC Vive Pro-Eye (https://vr.tobii.com/integrations/htc-vive-pro-eye/). From Tobii's API, we were able to collect data on pupil position, pupil diameter, gaze position, and gaze direction from participants as they worked on the CL tasks.

To minimize the use of noisy pupil data in our dataset, we devised a filtering mechanism based on the assumption that left and right eye pupil dilation should be highly correlated. First, a running average of the difference between left pupil dilation and right pupil dilation is calculated over a 120 sample (1 second) rolling window. Next, the standard deviation of the window is calculated. If the standard deviation exceeds a threshold, the window is considered too noisy. We empirically selected the threshold as 0.3 since a higher threshold compromises the data quality and a lower threshold leads to a scarcity of usable data. The percentage of noisy windows in a session is calculated by dividing the number of noisy windows across the total number of windows. Then if the percentage of noisy windows exceeds 30% of the overall data, the entire data associated with the subject is excluded from training. All of the data included in this dataset has passed this signal quality filter.

We focused primarily on pupillometry and rapid movements of the eye between points (i.e., saccades) because research suggests that these indicators are particularly sensitive to changes in cognitive load and relatively context independent [20]. Gaze data, on the other hand, is determined largely by the content presented (i.e., it is situation specific) and is not, itself, a reliable indicator of cognitive load. The Tobii output signals can be regarded as a multivariate time series that captures the values of the indicators over time. Table 2 includes a full list of features and details the sensor and the feature family (e.g., pupil features, fixation features) of each feature. We used the multivariate time series (MTS) of six variables, three for each eye (pupil diameter, pupil position on the x-axis, and pupil position on the y-axis). The procedure we used to process these signals and extract features is outlined below.

To enable real-time prediction of cognitive load, six-dimensional raw data from the Tobii sensor was segmented using a sliding window approach. For each window, the minimum, maximum, mean, and standard deviation of each dimension were calculated and updated to correspond with the current timestamp. These summary statistics from previous windows were utilized for real-time rolling normalization. This data processing step mitigates individual variations at the raw data level, allowing for consistent thresholding in detecting high-level features such as blinks, saccades, and fixations.

The signal from eye tracker also captured blinks, which introduces unwanted interference. To detect and remove blinks, we used guidance and techniques developed by Mathot, et al. [34] to identify and reconstruct the signal loss from closing the eyes during blinks. To start, we calculated standard deviation to normalize pupil diameter signals in rolling windows. If, during a given time interval, the standard deviation of pupil diameter was greater than a threshold of 0.4, we marked these as blink events. Once a blink event was identified, we calculated the longest blink duration, mean blink duration and blink rate (Hz) and used these as our blink features. We finally reconstructed the pupil data during a blink by removing the blink events and filling in the missing data with interpolation.

Next, we captured the influence of cognitive load on pupil dilation. After blink removal and data interpolation, pupil diameter

Table 2: Full list of features organized by sensor and feature family

Sensor	Feature Family	Feature	Unit
Pupilometry	Pupil Size	Pupil Diameter(mean)	mm
Pupilometry	Pupil Size	Pupil Diameter(std)	mm
Eye Tracking	Blink	Blink Rate	Hz
Eye Tracking	Blink	Longest Blink Duration	sec
Eye Tracking	Blink	Blink Depth	sec
Eye Tracking	Saccade	Number of Saccades (mean)	num./sec
Eye Tracking	Saccade	Number of Saccades (std)	num./sec
Eye Tracking	Saccade	Path length	degree
Eye Tracking	Saccade	Duration (mean)	sec
Eye Tracking	Saccade	Duration (std)	sec
Eye Tracking	Saccade	Saccade length (longest)	sec
Eye Tracking	Saccade	Saccade length (sum)	sec
Eye Tracking	Saccade	Saccade length (mean)	sec
Eye Tracking	Saccade	Speed (mean)	sec
Eye Tracking	Saccade	Speed (std)	sec
Eye Tracking	Saccade	Speed (max)	sec
Eye Tracking	Saccade	Rate	Hz
Eye Tracking	Fixation	Duration (longest)	sec
Eye Tracking	Fixation	Duration (sum)	sec
Eye Tracking	Fixation	Duration (mean)	sec
Eye Tracking	Fixation	Duration (std)	sec
Eye Tracking	Fixation Saccade Ratio	Rate	-
PPG	Heart Rate	Interbeat Interval	sec
PPG	Heart Rate	Standard Deviation of Normal Beats	sec
PPG	Heart Rate Variability	Standard Deviation of Successive Differences	sec
PPG	Heart Rate Variability	Root Mean Square of Successive Differences	sec
PPG	Heart Rate Variability	Avg. Power Spectral Density [0.0-0.2] (mean)	-
PPG	Heart Rate Variability	Avg. Power Spectral Density [0.2-0.4] (mean)	-
PPG	Heart Rate Variability	Avg. Power Spectral Density [0.4-0.6] (mean)	-
PPG	Heart Rate Variability	Avg. Power Spectral Density [0.6-0.8] (mean)	-
PPG	Heart Rate Variability	Avg. Power Spectral Density [1-2] (mean)	-

features were extracted by calculating the mean and standard deviation of the pupil diameter data.

We further capture saccade information from the eye data. Saccades are characterized by a quick, simultaneous movement of both eyes in the same direction and can be detected by calculating rapid changes in gaze direction. To calculate saccades in our data, we estimated the speed of gaze movement using pupil position data (e.g. left pupil position in tracking area x, and left pupil position in tracking area y). Similar to blink detection, we started by selecting time intervals when the speed of gaze movement exceeded a threshold. For all saccade events within the data buffer, we calculate statistical features of each saccade, such as number of saccades within the buffer, longest saccade, mean and standard deviation of the duration of saccades (second), and saccade rate (Hz). We also calculate statistical features that describe the speed of eye movement, such as maximum speed, mean and standard deviation of gaze movement speed, gaze movement path length, total saccade duration (second), and saccade rate (Hz).

Finally, we extract fixation events by selecting time intervals when changes in gaze direction is below the threshold. For all fixation events within the data buffer, we calculate statistical features, such as number of fixations, longest fixation, and mean and standard deviation of fixation durations (second), and fixation rate (Hz). We also calculate different ratios of saccade features to fixation features, such as the ratio of fixation rate over saccade rate, ratio of longest fixation duration over longest saccade duration, ratio of sum fixation duration over sum saccade duration, and ratio of mean fixation duration over mean saccade duration. The global minimum, maximum, mean, and standard deviation values of all the extracted features for each individual were tracked and updated from the start of recording until the end of the recording. All extracted features were normalized using the updated global statistics.

4.2 Pulse Plethysmography (PPG) Data

PPG is a light-based sensor that can measure heart activity by detecting changes in blood flow at the location where the sensor is applied to the skin (e.g., blood oxygenation and volume). In this data collection experiment, we used a Bitalino (r)evolution (http://bitalino.com) sensor board, with a PPG sensor applied to the fingertip of participants. PPG sensors are a reliable and relatively non-invasive way to estimate cardiac changes but the signal is particularly prone to artifacts and other sources of noise. For example, good contact between the sensor and the skin is necessary for reliable readings, making PPG sensitive to changes in movement. More complex sources of noise can occur because PPG measures the heart activity from pulses recorded in the periphery (arms, wrists, foreheads, etc.) leading to variability in the time it takes for the pulse to travel to the recording sight (called pulse transit time, or PTT). PTT can differ depending on predictable internal characteristics like a participants' height, age, or the health of their vasculature but PTT also varies with increased (or decreased) physiological arousal in complex, difficult to measure ways. To account for these sources of variability in the PPG signal, we developed a six-step processing algorithm to extract features known to be related to tracking and estimating cognitive load.

To filter noise and capture the signal with cardiac information, we used a bandpass FIR (finite impulse response) filter with a lower cutoff frequency of 0.5Hz and an upper cutoff frequency of 5Hz. Next, we used time-series decomposition tools for our analysis. In our experiments, removing trend from the PPG signal helped in better estimation of heart-rate features. We deconstructed the signal using seasonal decomposition, which decomposes the signal based on rates of change. This allowed us to remove slower moving component series in the PPG signal, returning a detrended PPG signal. The seasonal decomposition was performed using the statsmodels package in python (statsmodels.org). We then detected peaks in the detrended signal for calculating heart-rate features.

The prominent (Systolic) peaks in the PPG signal correspond to heartbeats and the proper detection of peaks is the core to calculating accurate heartrate features. As a first pass, our peak detection algorithm used the find_peaks algorithm in SciPy to detect peaks in the signal. However, the algorithm also detects secondary (diastolic) peaks, and other peaks resulting from noise. Elgendi [16] demonstrated that skewness (a measure of symmetry) of the photoplethysmogram (PPG) signals is the optimal method for determining the quality of the signal and can be formalized into a signal quality index (SQI). We used this SQI to further refine the peak de-

tection. Characteristic PPG signals have a skewed shape that can be attributed to the systolic and diastolic peaks of the signal. Because the systolic peak always occurs prior to the diastolic peak, in a tight time range around the systolic peak (e.g. +/- 160ms), we expect no influence of the signal from the diastolic peak, and therefore the signal should be relatively symmetric and skewness relatively low. If the skewness of the systolic peak exceeds this threshold, we filter the peak from the list because we assume that it is a false peak, most likely a diastolic peak misclassified as a systolic peak. Next, we consider an entire peak to peak interval. In this instance, both the systolic and diastolic peaks should be influencing the signal, causing a positive skewness (because the systolic peak should be larger than the diastolic peak). If the skewness for the RR interval is negative, the signal quality is low, and the RR interval is filtered from the list.

Despite the previous filtering steps, it is still possible for the diastolic peak to be misidentified as a systolic peak. To remove these artifacts, we deploy an algorithm similar to Lipponen and Tarvainen [30, 57, 58] that examines the RR series, which is a time series of successive RR intervals. We compare the current interbeat interval (IBI) to the 3 most recent IBIs. If the interbeat interval is within a 150ms threshold of any of the 3 most recent IBIs, it is considered valid. If not, then we attempt to combine the current IBI with the next IBI and recheck to see if the new combined IBI is within the 150ms threshold.

Finally, with a full list of valid IBI intervals, it is trivial to compute the heart rate, and pulse rate variability. Our features use a sliding window to calculate both heart rate and pulse rate variability, where heart rate = 1/avg (IBI) and pulse rate variability can be characterized as the successive differences of normal beats (SDNB), and root mean square of successive differences (RMSSD) over the time window. Note that we use the term pulse rate variability instead of heart rate variability since we are measuring heart activity through the circulatory system, which layers a variable pulse transit time (PTT) to the pulse rate variability statistic. We further calculate power spectral density (PSD) features across the frequency bands [0.0-0.2] Hz, [0.2-0.4] Hz, [0.4-0.6] Hz, [0.6-0.8] Hz, [1-2] Hz, to capture information in the frequency domain. The PSD features were calculated using Welch method [16].

4.3 Cognitive Load Normalization for 'Ground Truth' Labels

One of the biggest challenges for predicting cognitive load in a context independent way is that it is very difficult to ascertain "ground truth" cognitive load levels. This is a challenge for data quality, especially label quality, because the machine learning model is trained on the labeled data. The closer the labels are to the "true" measure of interest, the more likely that the model will perform well on unlabeled data. We have implemented a multi-pronged labeling approach (e.g., [5]) that combines indices of task difficulty and subjective ratings of mental load to create a labeling paradigm that most reliably represents the true level of cognitive load experienced by individuals in a given moment. Once we validated the subjective ratings, we used them in combination with task difficulty to label cognitive load for each individual trial.

Using z-score normalization, the resulting data distribution is centered at 0 with 95% of the samples between [-2,2]. Our desired score is between [0,1], so we transform the z-score using the following formula.

$$I_n = I_z/4 + 0.5, \tag{1}$$

where I_n represents normalized subjective rating, and I_z is the z score rating. Although technically only 95% of the resulting samples are between the desired range of [0,1], this is sufficient for our needs.

In our stimulus regime, task difficulty increases are additive. Easier tasks are entirely contained within the more difficult tasks.

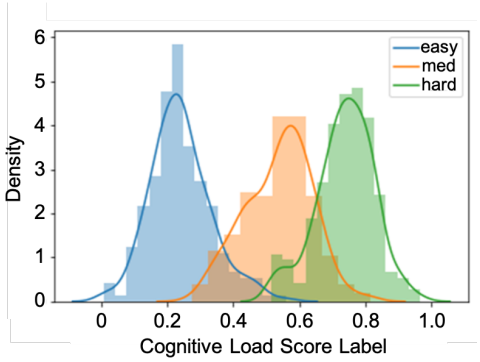

Figure 3: Density of labels separated by task difficulty

Table 3: Additive Difficulty Weighting by Task

Task Description	Normalized Difficulty Rating
Task 1: Visual Vigilance	0.25
Task 2: Visual Vigilance + Arithmetic	0.50
Task 3: Visual Vigilance + Arithmetic + Audio Vigilance	0.75

To represent this, we assigned each task a weighted difficulty rating for each task (See Table 3 for values by task).

We then combine individual subjective rating with task averages using a weighted sum

$$R_n = T_n * 0.5 + I_n * 0.5, \tag{2}$$

where R_n, T_n, and I_n represent normalized subjective rating, population wide normalized task rating, and individual normalized rating. Fig. 3 illustrates the resulting distribution of labels for the different cognitive load scores for each task difficulty condition. The dataset labels are exclusively derived from the study itself, based on participants' subjective ratings and stimulus sampling. No expert labels or third-party annotations are included, and the labels solely reflect the participants' first-person ratings of their subjective experiences. Because we only have one subjective rating for the entire task, we assume that the cognitive load score label is constant for the duration of the task.

5 Modeling

5.1 Model Selection

We trained a deep learning model to predict cognitive load as a continuous value between 0 and 1, representing the lowest and highest reported cognitive loads. To handle multi-modal data, we use a deep neural network instead of traditional methods like SVM. Deep learning allows for feature and decision fusion, providing more integrated representations, and supports end-to-end training, eliminating the need for separate feature extraction. Our architecture also accommodates both regression and classification tasks, enhancing versatility and efficiency.

Our study has three types of tasks, *easy*, *medium*, and *hard*, which are expected to induce *low*, *medium*, and *high* cognitive loads respectively. However, individual cognitive capacities cause subjective load experiences to vary. Fig. 3 shows that the reported cognitive loads approximately follow a Gaussian distribution with respect to each type of task and the three distributions overlap with each other. Collectively, the reported cognitive loads follow a distribution with multiple modes. Regression with a multi-modal distribution involves identifying the mode and regressing the offset from

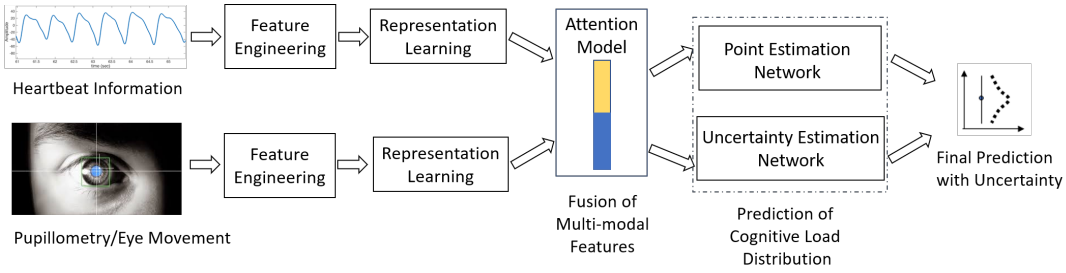

Figure 4: Machine learning model architecture for cognitive load prediction from signals

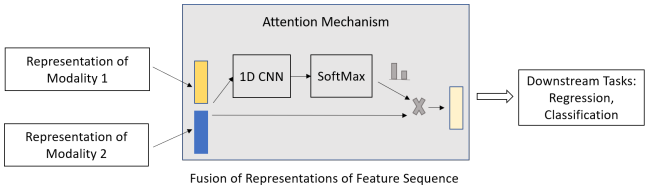

Figure 5: Illustration of high-level representation learning through sensor fusion

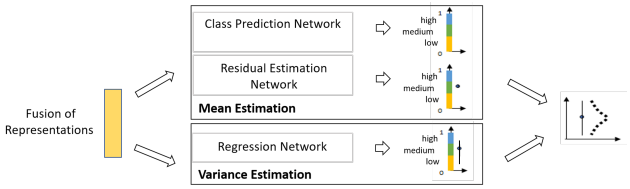

Figure 6: Depicts architecture from fused representation through cognitive load distribution prediction to final prediction with uncertainty

the mode center. Inspired by Multi-Bin loss regression [38], our approach predicts one of the three cognitive load levels and estimates the residual relative to the average load of each level. Additionally, we model prediction uncertainty by predicting a cognitive load distribution, such as a Gaussian distribution with mean and standard deviation.

5.2 Training Data Preparation

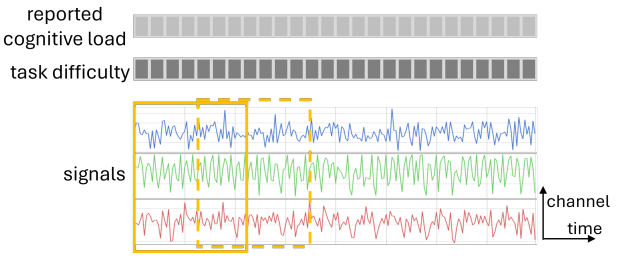

Figure 7: Features sequence segments and corresponding labels. We use three feature sequences to illustrate an arbitrary number of n features extracted from raw signals.

To train our regression model, we constructed pairs of input feature sequences and the corresponding cognitive load labels. Using the methods articulated in Section 4, we can convert raw signals to temporal feature sequences. Fig. 7 shows how the input features sequence $\mathbf{x}$, and the output, subject cognitive load rating y_r and task difficulties y_c, are paired.

The input feature sequences $\mathbf{x}$ are temporal multivariate data derived from physiological signals, including pupilometry, eye-tracking, and pulse plethysmography (PPG). The input $\mathbf{x} \in \mathbb{R}^{t \times d}$ represents a sequence of d-dimensional features of length t, where t is the time dimension and d represents the feature dimension. The outputs include a real-valued normalized subjective cognitive load rating $y_r \in [0,1]$ and a discrete task difficulty $y_c \in \{0,1,2\}$ representing the objective task difficulties of *easy*, *medium*, and *hard*, respectively. These labels were derived from a combination of subjective ratings and normalized task difficulties, as described in Section 4. For data from each experiment trial, we begin extracting data and label pairs only after the initial few seconds to eliminate any unstable phase at the start of the trial. The input features and labels were synchronized using a sliding window approach to ensure temporal alignment. A typical sliding window size is 120 samples, while the skip step for the rolling window is significantly smaller, typically just 1 sample. This enables cognitive load predictions to be made at each skip step.

5.3 General Model Architecture

The model architecture is shown in Fig. 4. Feature extraction and representation learning use convolutional neural networks (CNNs) to process physiological signals and capture temporal patterns. Attention mechanism weighs the importance of specific time steps or features. Each CNN branch includes two 1D convolutional layers with kernel size 3, followed by batch normalization, ReLU activation, and max pooling with kernel size 2. These layers extract and downsample relevant features from each input modality. Global average pooling then condenses the feature maps into fixed-size 64-dimensional vectors, enabling sensor fusion.

The sensor fusion step (Fig. 5) combines feature vectors from two CNN branches into a unified 128-dimensional representation, refined by a dual-branch attention mechanism. This mechanism uses channel and spatial attention to emphasize key features and reduce noise sensitivity. Channel attention generates weights by pooling across the temporal dimension and processing through an MLP, while spatial attention focuses on critical temporal regions. These attention modules enable dynamic feature weighting, improving robustness and accuracy. Fig. 5 illustrates how this process facilitates seamless integration across modalities.

The fused representation is used to estimate a probabilistic distribution of the possible prediction values, as illustrated in Fig. 6. We characterize this distribution as a Gaussian distribution, whose mean and standard deviation parameters can be modeled as two neural networks. The final layer of the model outputs two values $\mu(\mathbf{x})$ and $\delta(\mathbf{x})$, which correspond to the predicted mean value and standard variation. By treating the ground truth cognitive load value as a sample from the Gaussian distribution parameterized by $\mu(\mathbf{x})$ and $\delta(\mathbf{x})$, following the idea in [39], we use a stochastic gradient descent algorithm to train the machine learning model end to end

by minimizing the negative log-likelihood as in Equation 3:

$$-log p_\theta(y_n \mid \mathbf{x}_n) = \frac{log\sigma_\theta^2(\mathbf{x}_n)}{2} + \frac{(y_n - \mu_\theta(\mathbf{x}_n))^2}{2\sigma_\theta^2(\mathbf{x}_n)^2} + constant \quad (3)$$

The proposed model offers several advantages. It achieves real-time inference by leveraging efficient feature extraction and sliding window normalization. The probabilistic framework not only predicts cognitive load but also estimates confidence intervals, aiding decision-making under uncertainty. Robust multi-modal fusion integrates complementary signals, enhancing prediction accuracy and mitigating noise impacts. Furthermore, the architecture is adaptable to diverse datasets and application scenarios, ensuring broad applicability in VR contexts.

The model was trained using a single NVIDIA GeForce GTX 1080Ti GPU with PyTorch as the deep learning framework. The data was processed in batches of 32 samples. The Adam optimizer was employed with a learning rate of 0.0001, and a learning rate decay was applied every 10 epochs using a StepLR scheduler with a decay factor (gamma) of 0.3. Training was conducted for 12 epochs with early stopping and model checkpointing enabled. For deployment, the trained model was converted to ONNX format.

6 Results and Discussion

6.1 Result

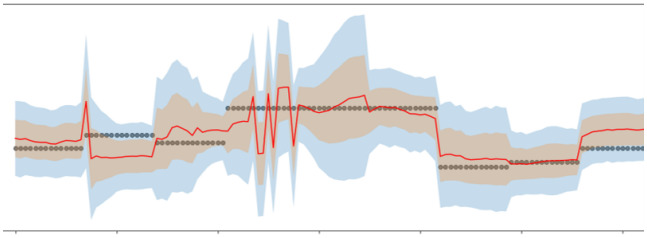

Figure 8: Cognitive load prediction for a test user through time (red). The black dots represent the 'ground truth' cognitive load score, which is determined for each task by Eq. 2 in section 4.3. Based on the prediction of cognitive load variance, the lower and upper bounds of cognitive load prediction are illustrated as the brown region (one standard deviation of the variance) and the blue region (two times standard deviation of the variance).

We evaluate the efficacy of our proposed method by training the model on a comprehensive dataset comprising physiological measurements from 638 individuals and subsequently testing its generalizability on an independent dataset from 100 individuals, the latter of which has been made publicly available to facilitate reproducibility and further research. The testing dataset includes diverse physiological signals, specifically pupilometry/eye tracking data and photoplethysmography (PPG) signals, which serve as input modalities for cognitive load and task difficulty level prediction. The testing data are made available to the public. A one dimensional CNN model is used to learn representations from temporal feature sequences and multiple layer perception is used to predict the cognitive load levels, the residual within each level, and the uncertainty variation. The model is trained end-to-end using a Pytorch implementation of stochastic gradient descent with mini-batches of 32 data samples.

Table 4 presents the evaluation results, demonstrating that our method achieves a mean absolute error (MAE) of 0.11 in predicting cognitive load by effectively fusing information from both pupilometry/eye tracking data and the PPG signal. This performance indicates a high degree of precision in cognitive load estimation,

Table 4: Inference Engine Evaluation using Multimodal Inputs

	Fusion Model	Tobii Model	PPG Model
Cognitive Load Prediction (Mean Absolute Error)	0.1105	0.1131	0.1755
Task Difficulties Prediction (Accuracy)	78.82%	73.45%	53.42%

underscored by the model's ability to integrate multi-modal physiological data.

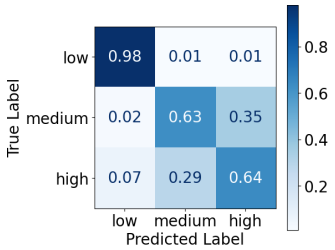

Figure 9: Confusion Matrix for Task Difficulty Level Prediction

Furthermore, the model exhibits a "by-product" capability in predicting task difficulty levels, attaining a best classification accuracy of 78.82% using the fusion approach. This dual-task functionality not only enhances the model's utility but also leverages shared feature representations to improve prediction accuracy across related tasks.

The confusion matrix (Fig. 9), based on one of our training results, give a view of the model's classification performance. These results indicate that the model distinguishes low task difficulty levels effectively, but often confuses medium and high difficulty levels. One possible explanation for this is that many participants' cognitive load may have become saturated with the medium difficulty stimuli, making it more difficult to differentiate from the high difficulty stimuli. However, we note that the model uses this task difficulty prediction internally and does not report it as part of the predicted cognitive load score. The resulting cognitive load score performed reasonably well in terms of MSE, despite this confusion rate because of categorical overlap that can be observed in Figure 3. In other words, the scores overlap such that a low mental workload rated experience in a high difficulty stimulus task scores the same as a high mental workload rated experience in a medium difficulty stimulus task.

Figure 8 illustrates the cognitive load predictions for a representative user during the testing phase. The majority of predictions fall within one standard deviation of the estimated uncertainty, underscoring the model's reliability and the effectiveness of its uncertainty quantification. This uncertainty estimation is derived from the regression head's ability to predict both the mean cognitive load and its associated variance, providing valuable insights into the confidence of each prediction.

6.2 Early Wins

Our efforts in demonstrating the use of real-time cognitive load inference across various scenarios, such as learning and training in VR applications, extend beyond mere prototyping. Its success during the incubation phase, where off-the-shelf hardware was employed, has led to the creation of HP Reverb G2 Omnicept Edition commercial VR headset. The cognitive load models have since been refined using data collected through the HP Reverb G2 Omnicept Edition headsets, resulting in a more robust and integrated

solution.

Our cognitive load inference model is now actively deployed and being utilized in diverse real-world applications by both researchers and developers. For example, the company Ovation VR is using the cognitive load inference to implement assessment and adaptive learning for public speaking training. Another company called PIXO uses the measure to provide real-time feedback for virtual reality training across industries such as construction, manufacturing and public safety. The cognitive load inference is also widely used by academic researchers to study areas such as virtual commerce in VR [60], VR gaming [36], and virtual assistant in driving [61].

6.3 Implications for Designers

Cognitive load can improve training as a predictor of performance (Fig. 10). Yerkes-Dodson [59] describes an inverted U-curve linking cognitive load and performance. Excessive work load leads to cognitive overload, errors, and inefficiency, while minimal load results in disengagement and reduced focus. The optimal "Goldilocks zone" balances challenge and engagement for peak performance.

Cognitive load can also improve training as a predictor of expertise, particularly when combined with measures of performance. Early training shows high load and low performance, which improves as proficiency grows. Mastery occurs when performance peaks and load stabilizes in the optimal range. Tracking this relationship helps streamline training by focusing on areas needing improvement.

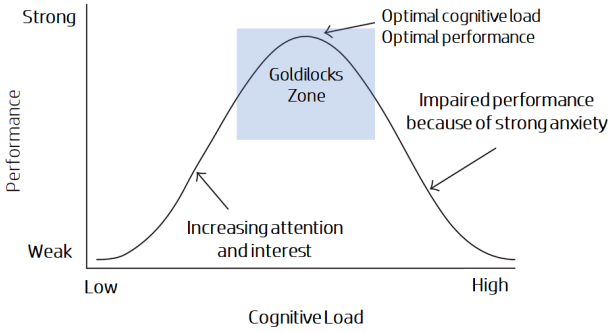

Figure 10: Cognitive Load as a predictor of performance. Performance is optimal when cognitive load is in the Goldilocks zone.

6.4 Limitations and Future Work

The cognitive load inference engine can be improved over time by accounting for some of the limitations in this dataset. Since this open dataset was collected, we have expanded our data collection activities to include data from several additional locations across the world, including data from participants in Africa, Asia, and North America. The pervasive bias in AI can be further mitigated and one way to do that is by collecting data from as wide a sample of the general population as possible. Our current multi-tasking paradigm varies in both difficulty and modality but we have not, yet, tested our models in other types of cognitive load contexts (e.g., with time or social evaluative pressure, with motor load, with continuously increasing difficulty). Including new manipulations of load will increase the inference engine capability of generalization. This work can be expanded by exploiting particular features of virtual environments as well.

The goal of this paper is to develop a person-independent cognitive load inference model that requires no calibration so that it can be easily integrated into VR applications. In practice, significant inter-individual variability in physiological responses to workload makes it challenging to apply a universal set of cognitive load measures across all individuals. One next step for our work could involve developing a cognitive load model fine-tuned to individuals or specific groups based on personal traits such as age, gender, or educational background, which could significantly enhance the system's ability to detect cognitive load. Additionally, if sufficient training data is available for an individual, the model could be fine-tuned to accurately reflect that person's unique characteristics. However, implementing such customizations to address individual differences in cognitive load assessment would require additional data collection, design, and engineering considerations for effective integration into VR applications.

Additionally, having fine-grained ground truth labels is critical for achieving real-time cognitive load inference. Relying on a single measure, such as self-reports, task difficulty, or task performance, as the ground truth for cognitive load presents significant challenges. An innovative approach that cross-references these measures could provide a more robust and reliable method for labeling cognitive load. However, this will require more fundamental research in cognitive load theory and measurement.

Cognitive load inference using physiological sensors holds great promise but faces practical challenges in real-world applications. For instance, dynamic physical activities, such as sports, often involve rapid and unpredictable movements that can introduce noise and artifacts into physiological signals like photoplethysmography (PPG) or eye-tracking data. Mitigating these effects requires advanced signal processing techniques, such as adaptive filtering and motion-compensation algorithms. Lighting interference within VR headsets can also impact the accuracy of PPG or eye-tracking systems. Addressing this issue necessitates the development of robust sensors with enhanced shielding against ambient light and algorithms designed to compensate for lighting variations effectively.

While we have developed a neural network for cognitive load inference, future work should include a comprehensive performance analysis comparing traditional methods and emerging architectures, focusing on both cognitive load prediction and uncertainty quantification. This will help establish a robust baseline for machine learning-based cognitive load inference research.

After all, we hope this paper and dataset serve as valuable resources for technologists, researchers, and machine learning scientists, fostering further advancements and discussions within the scientific community. We encourage researchers of all stripes to test, validate, and reach out to our team with questions, thoughts, or insights.

6.5 Conclusion

Researchers are increasingly studying VR for training and learning. One of the more robust findings from this recent research is that the very high presence that makes VR so engaging and compelling also causes cognitive load, and interferes with learning (See Makransky & Mayer [32] for a recent review, or Bailey et al. [6] for an early example). This is a frustrating combination – when VR is at its best and users are highly engaged perceptually and psychologically, the medium can interfere with learning. Our findings provide a possible solution to this problem by detecting cognitive load dynamically one can modulate presence to ensure learning proceeds optimally.

In this paper, we present a pioneering large-scale study aimed at devising a machine learning model that can reliably predict real-time mental effort in VR. We collected physiological, self-report, and task data from more than seven hundred participants who completed mentally demanding tasks of increasing difficulty. We developed an innovative data filtering and labeling pipeline, and trained a multi-modal, fusion model that predicted real-time cognitive load with low error. We are also making available our test dataset to advance the research community's resources. This comprehensive dataset contains multi-modal data from 100 distinct users, serving as a pivotal resource for future investigations in the domain.

References

[1] S. Afzal and P. Robinson. Natural affect data collection and annotation in a learning context. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–7, 2009. doi: 10.1109/ACII.2009.5349537 2

[2] M. U. Ahmed, S. Begum, R. Gestlöf, H. Rahman, and J. Sörman. Machine learning for cognitive load classification–a case study on contact-free approach. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 31–42, 2020. 2

[3] P. Ayres. Impact of reducing intrinsic cognitive load on learning in a mathematical domain. *Applied Cognitive Psychology*, 20:287–298, 2006. 2

[4] A. D. Baddeley and G. Hitch. Working memory. vol. 8 of *Psychology of Learning and Motivation*, pp. 47–89. Academic Press, 1974. doi: 10.1016/S0079-7421(08)60452-1 1, 2

[5] J. N. Bailenson, K. Swinth, C. Hoyt, S. Persky, A. Dimov, and J. Blascovich. The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoper. Virtual Environ.*, 14(4):379–393, Aug. 2005. doi: 10.1162/105474605774785235 6

[6] J. Bailey, J. N. Bailenson, A. S. Won, J. Flora, and K. C. Armel. Presence and memory: immersive virtual reality effects on cued recall. In *Proceedings of the International Society for Presence Research Annual Conference*, vol. 10, pp. 24–26, 2012. 9

[7] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20:1 – 68, 2019. 2

[8] M. Bartels and S. P. Marshall. Measuring cognitive workload across different eye tracking hardware platforms. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, p. 161–164. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2168556.2168582 3, 4

[9] P. Biswas and G. Prabhakar. Detecting drivers' cognitive load from saccadic intrusion. *Transportation Research Part F-traffic Psychology and Behaviour*, 54:63–78, 2018. 2

[10] H. Boril, S. O. Sadjadi, and J. Hansen. Utdrive: Emotion and cognitive load classification for in-vehicle scenarios. 2011. 2

[11] S. Chen, J. Epps, and F. Chen. A comparison of four methods for cognitive load measurement. *Proceedings of the 23rd Australian Computer-Human Interaction Conference, OzCHI 2011*, 11 2011. doi: 10.1145/2071536.2071547 1

[12] P. Chwalek, D. Ramsay, and J. A. Paradiso. Captivates: A smart eyeglass platform for across-context physiological measurement. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(3), sep 2021. doi: 10.1145/3478079 2

[13] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. In *Computer Vision and Image Understanding*, pp. 160–187, 2003. 2

[14] N. Cowan. The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, 19(1):51–57, 2010. PMID: 20445769. doi: 10.1177/0963721409359277 1

[15] R. D. Dias, M. A. Zenati, R. Stevens, J. M. Gabany, and S. J. Yule. Physiological synchronization and entropy as measures of team cognitive load. *Journal of Biomedical Informatics*, 96:103250, 2019. doi: 10.1016/j.jbi.2019.103250 2

[16] M. Elgendi. Optimal signal quality index for photoplethysmogram signals. *Bioengineering*, 3(4), 2016. doi: 10.3390/bioengineering3040021 5, 6

[17] R. Engle and M. Kane. Executive attention, working memory capacity, and a two-factor theory of cognitive control. *Psychology of Learning and Motivation*, 44:145–199, 2003. 1, 2

[18] E. Ferreira, D. Ferreira, S. Kim, P. Siirtola, J. Röning, J. F. Forlizzi, and A. K. Dey. Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. In *2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*, pp. 39–48, 2014. doi: 10.1109/CCMB.2014.7020692 2

[19] L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman. *Cognitive Load Estimation in the Wild*, p. 1–9. Association for Computing Machinery, New York, NY, USA, 2018. 2

[20] M. Gjoreski, T. Kolenik, T. Knez, M. Luštrek, M. Gams, H. Gjoreski, and V. Pejović. Datasets for cognitive load inference using wearable sensors and psychological traits. *Applied Sciences*, 10(11), 2020. doi: 10.3390/app10113843 2, 4

[21] E. Haapalainen Ferreira, S. Kim, J. Forlizzi, and A. Dey. Psychophysiological measures for assessing cognitive load. pp. 301–310, 12 2010. doi: 10.1145/1864349.1864395 2

[22] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, eds., *Human Mental Workload*, vol. 52 of *Advances in Psychology*, pp. 139–183. North-Holland, 1988. doi: 10.1016/S0166-4115(08)62386-9 4

[23] J. A. Healey and R. W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *Trans. Intell. Transport. Sys.*, 6(2):156–166, June 2005. doi: 10.1109/TITS.2005.848368 2

[24] M. Iskander. Burnout, cognitive overload, and metacognition in medicine. *Medical Science Educator*, 29, 11 2018. doi: 10.1007/s40670-018-00654-5 2

[25] M. R. Islam, S. Barua, M. U. Ahmed, S. Begum, and G. D. Flumeri. Deep learning for automatic eeg feature extraction: An application in drivers' mental workload classification. In *H-WORKLOAD*, 2019. 2

[26] P. Jercic, C. Sennersten, and C. Lindley. Modeling cognitive load and physiological arousal through pupil diameter and heart rate. *Multimedia Tools and Applications*, 79:3145–3159, 2018. 2

[27] A. Jimenez-Molina, C. Retamal, and H. Lira. Using psychophysiological sensors to assess mental workload during web browsing. *Sensors*, 18(2), 2018. doi: 10.3390/s18020458 2

[28] T. I. Kaluarachchi, S. Sapkota, J. Taradel, A. Thevenon, D. J. Matthies, and S. Nanayakkara. Eyeknowyou: A diy toolkit to support monitoring cognitive load and actual screen time using a head-mounted webcam. In *Adjunct Publication of the 23rd International Conference on Mobile Human-Computer Interaction*, MobileHCI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3447527.3474850 1

[29] J. Y. Lee, N. de Jong, J. Donkers, H. Jarodzka, and J. J. G. van Merriënboer. Measuring cognitive load in virtual reality training via pupillometry. *IEEE Trans. Learn. Technol.*, pp. 1–7, 2023. 1

[30] J. A. Lipponen and M. P. Tarvainen. A robust algorithm for heart rate variability time series artefact correction using novel beat classification. *Journal of Medical Engineering & Technology*, 43(3):173–181, 2019. PMID: 31314618. doi: 10.1080/03091902.2019.1640306 6

[31] K. P. Madore, A. M. Khazenzon, C. W. Backes, J. Jiang, M. Uncapher, A. Norcia, and A. Wagner. Memory failure predicted by attention lapsing and media multitasking. *Nature*, 587:87 – 91, 2020. 2

[32] G. Makransky and R. E. Mayer. Benefits of taking a virtual field trip in immersive virtual reality: Evidence for the immersion principle in multimedia learning. *Educational Psychology Review*, 34(3):1771–1798, Sep 2022. doi: 10.1007/s10648-022-09675-4 9

[33] G. Makransky and G. B. Petersen. The cognitive affective model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educ. Psychol. Rev.*, 33(3):937–958, Sept. 2021. 1

[34] S. Mathôt. A simple way to reconstruct pupil size during eye blinks. *FigShare*, 04 2013. doi: 10.6084/m9.figshare.688001 4

[35] R. McKendrick, B. Feest, A. Harwood, and B. Falcone. Theories and methods for labeling cognitive workload: Classification and transfer learning. *Frontiers in Human Neuroscience*, 13:295, 2019. doi: 10.3389/fnhum.2019.00295 2

[36] D. Mendez-Encinas, A. Sujar, S. Bayona, and D. Delgado-Gomez. Attention and impulsivity assessment using virtual reality games. *Sci. Rep.*, 13(1):13689, Aug. 2023. 9

[37] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97, March 1956. 1

[38] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka. 3d bound-

ing box estimation using deep learning and geometry. *CoRR*, abs/1612.00496, 2016. 7

[39] D. Nix and A. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, vol. 1, pp. 55–60 vol.1, 1994. doi: 10.1109/ICNN.1994.374138 7

[40] N. Nourbakhsh, Y. Wang, and F. Chen. Gsr and blink features for cognitive load classification. In *INTERACT*, 2013. 2

[41] K. Oberauer, H.-M. Süß, O. Wilhelm, and N. Sander. *Individual Differences in Working Memory Capacity and Reasoning Ability*, pp. 49–75. 03 2008. doi: 10.1093/acprof:oso/9780195168648.003.0003 1, 2

[42] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. V. Gerven. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1):63–71, 2003. 1, 2

[43] K. Qi, D. Borland, N. L. Williams, E. Jackson, J. Minogue, and T. C. Peck. Augmenting physics education with haptic and visual feedback. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 439–443, 2020. doi: 10.1109/VRW50115.2020.00093 1

[44] S. Ransbotham. The subtle sources of sampling bias hiding in your data. *MIT Sloan Management Review*, 59(1), 2017. 1

[45] C. Ranti, W. Jones, A. Klin, and S. Shultz. Blink rate patterns provide a reliable measure of individual engagement with scene content. *Scientific Reports*, 10, 2020. 2

[46] W. L. Romine, N. L. Schroeder, J. Graft, F. Yang, R. Sadeghi, M. Zabihimayvan, D. Kadariya, and T. Banerjee. Using machine learning to train a wearable device for measuring students' cognitive load during problem-solving activities based on electrodermal activity, body temperature, and heart rate: Development of a cognitive load tracker for both personal and classroom use. *Sensors*, 20(17), 2020. doi: 10.3390/s20174833 2

[47] D. Roselli, J. Matthews, and N. Talagala. Managing bias in ai. pp. 539–544, 05 2019. doi: 10.1145/3308560.3317590 1

[48] A. Saha, V. Minz, S. Bonela, S. Sr, R. Chowdhury, and D. Samanta. *Classification of EEG Signals for Cognitive Load Estimation Using Deep Learning Architectures: 10th International Conference, IHCI 2018, Allahabad, India, December 7–9, 2018, Proceedings*, pp. 59–68. 12 2018. doi: 10.1007/978-3-030-04021-5_6 2

[49] P. Sarkar, K. Ross, A. J. Ruberto, D. Rodenbura, P. Hungler, and A. Etemad. 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)*, pp. 1–7, 2019. doi: 10.1109/ACII.2019.8925507 2

[50] D. Schlör, A. Zehe, K. Kobs, B. Veseli, F. Westermeier, L. Brübach, D. Roth, M. E. Latoschik, and A. Hotho. Improving sentiment analysis with biofeedback data. In P. Paggio, A. Gatt, and R. Klinger, eds., *Proceedings of LREC2020 Workshop "People in language, vision and the mind" (ONION2020)*, pp. 28–33. European Language Resources Association (ELRA), Marseille, France, May 2020. 1

[51] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):410–417, 2010. doi: 10.1109/TITB.2009.2036164 2

[52] T. M. Sezgin and P. Robinson. Affective video data collection using an automobile simulator. In A. Paiva, R. Prada, and R. W. Picard, eds., *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, vol. 4738 of *Lecture Notes in Computer Science*, pp. 770–771. Springer, 2007. doi: 10.1007/978-3-540-74889-2_91 2

[53] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. doi: 10.1109/T-AFFC.2011.25 1

[54] J. Su and S. Luz. *Predicting Cognitive Load Levels from Speech Data*, pp. 255–263. 01 2016. doi: 10.1007/978-3-319-28109-4_26 2

[55] J. Sweller. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4):295–312, 1994. doi: 10.1016/0959-4752(94)90003-5 2

[56] J. Sweller. Cognitive load theory. vol. 55 of *Psychology of Learning and Motivation*, pp. 37–76. Academic Press, 2011. doi: 10.1016/B978-0-12-387691-1.00002-8 1, 2

[57] M. Tarvainen, J. Niskanen, J. Lipponen, P. Ranta-aho, and P. Karjalainen. *Kubios HRV — A Software for Advanced Heart Rate Variability Analysis*, vol. 22, pp. 1022–1025. 01 2009. doi: 10.1007/978-3-540-89208-3_243 6

[58] M. P. Tarvainen, J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-aho, and P. A. Karjalainen. Kubios hrv – heart rate variability analysis software. *Computer Methods and Programs in Biomedicine*, 113(1):210–220, 2014. doi: 10.1016/j.cmpb.2013.07.024 6

[59] K. H. Teigen. Yerkes-dodson: A law for all seasons. *Theory & Psychology*, 4(4):525–547, 1994. doi: 10.1177/0959354394044004 1, 9

[60] T. Weiß and J. Pfeiffer. Consumer decisions in virtual commerce: Predict good help-timing based on cognitive load. *J. Neurosci. Psychol. Econ.*, 17(2):119–144, June 2024. 9

[61] S. Wen, S. Ping, J. Wang, H.-N. Liang, X. Xu, and Y. Yan. AdaptiveVoice: Cognitively adaptive voice interface for driving assistance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–18. ACM, New York, NY, USA, May 2024. 9

[62] D. Wu, J. Ouyang, N. Dai, M. Wu, H. Tan, H. Deng, Y. Fan, D. Wang, and Z. Jin. Deepbrain: Enabling fine-grained brain-robot interaction through human-centered learning of coarse eeg signals from low-cost devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(3), sep 2022. doi: 10.1145/3550334 2

[63] R. M. Yerkes and J. D. Dodson. The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology and Psychology*, 18:459–482, 1908. 1

[64] L. Zhang, J. Wade, D. Bian, J. Fan, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar. Cognitive load measurement in a virtual reality-based driving system for autism intervention. *IEEE Transactions on Affective Computing*, 8(2):176–189, 2017. doi: 10.1109/TAFFC.2016.2582490 2