2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) | 979-8-3315-1647-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISMAR62088.2024.00121

Mazed and Confused: A Dataset of Cybersickness, Working Memory, Mental Load, Physical Load, and Attention During a Real Walking Task in VR

Jyotirmay Nag Setu*
The University of Texas at San Antonio

Joshua M Le†
The University of Texas at San Antonio

Ripan Kumar Kundu‡
University of Missouri-Columbia

Barry Giesbrecht §
University of California, Santa Barbara

Tobias Höllerer ¶
University of California, Santa Barbara

Khaza Anuarul Hoque ‖
University of Missouri, Columbia

Kevin Desai **
The University of Texas at San Antonio

John Quarles ††
The University of Texas at San Antonio

Figure 1: a) Top-down view of a randomly generated virtual maze environment b) First-person point of view - a participant grabs the trophy to start another random maze task c) Participants physically walked in the real environment to enable virtual environment navigation

ABSTRACT

Virtual Reality (VR) is quickly establishing itself in various industries, including training, education, medicine, and entertainment, in which users are frequently required to carry out multiple complex cognitive and physical activities. However, the relationship between cognitive activities, physical activities, and familiar feelings of cybersickness is not well understood and thus can be unpredictable for developers. Researchers have previously provided labeled datasets for predicting cybersickness while users are stationary, but there have been few labeled datasets on cybersickness while users are physically walking. Moreover, it is unclear how walking while cybersick will affect cognitive load, even though room-scale interaction is typical in many VR games. Thus, from 39 participants, we collected head orientation, head position, eye tracking, images, physiological readings from external sensors, and the self-reported cybersickness severity, physical load, and mental load in VR. Throughout the data collection, participants navigated mazes via real walking and performed tasks challenging their attention and working memory. To demonstrate the dataset's utility, we conducted a case study of training classifiers in which we achieved 95% accuracy for cybersickness severity classification. The noteworthy performance of the straightforward classifiers makes this dataset ideal for future researchers to develop cybersickness detection and reduction models. To better understand the features that helped with classification, we performed SHAP(SHapley Additive exPlanations) analysis, highlighting the importance of eye tracking and physiological measures for cybersickness prediction while walking. This open dataset can allow future researchers to study the connection between cybersickness and cognitive loads and develop prediction models. This dataset will empower future VR developers to design efficient and effective Virtual Environments by improving cognitive load management and minimizing cybersickness.

Index Terms: virtual reality, real walking, cybersickness, cognitive load, attention, working memory, datasets, user studies, machine learning, deep learning, classification, explainable AI

*e-mail: jyotirmaynag.setu@utsa.edu
†e-mail: joshua.le@my.utsa.edu
‡e-mail: rkcgc@missouri.edu
§e-mail: giesbrecht@ucsb.edu
¶e-mail: holl@cs.ucsb.edu
‖e-mail: hoquek@missouri.edu
**e-mail: kevin.desai@utsa.edu
††e-mail: John.Quarles@utsa.edu

1 INTRODUCTION

Human-computer interaction now has new avenues to explore, thanks to the development of immersive virtual reality technology. From education, healthcare, and industrial training to gaming[8, 47,

979-8-3315-1647-5/24/$31.00 ©2024 IEEE
DOI 10.1109/ISMAR62088.2024.00121

20], these technologies can potentially change several industries. However, individuals may encounter physiological and psychological difficulties when they immerse themselves in these virtual experiences. In many popular VR experiences, players are required to perform multiple tasks concurrently, some of which are physical and some of which are mental [3, 65]. For example, consider the popular room-scale VR game Half Life: Alyx, in which players must battle enemies (physical), solve puzzles (mental), and navigate unfamiliar environments (physical and mental). For a developer, finding a balance between mental and physical demands in VR can be a daunting task that currently requires extensive playtesting. Unfortunately, there are few comprehensive datasets in the literature that can be used to help predict and combat physical and mental demands on VR users. To address this research gap, this paper provides a novel dataset intended to enable the assessment and prediction of cybersickness, working memory, mental load, physical load, and attention during room-scale VR experiences.

Few prior works have investigated the intersection of cybersickness, cognition, and real walking[60, 24]. For example, several cybersickness-related works have focused on collecting datasets for cybersickness prediction in VR [18, 16] while users were mostly stationary and did not consider the impact on cognition. Luong et al.[36] ran a large lab-in-the-field study with participants who used real-walking to navigate a 20x20 ballroom, but they failed to address the potential impact of cybersickness on cognition. In contrast to the prior work, our approach was to collect cybersickness and cognitive data while participants were physically walking.

Real walking in VR is becoming more commonplace due to off-the-shelf hardware support for room-scale locomotion, wide area navigation, and location-based VR, which can utilize a large amount of physical space[51]. Suma et al. [55] pointed out that in scenarios demanding swift and efficient navigation or travel mirroring real-world movements, genuine walking offers benefits over conventional joystick-driven virtual traversal methods. In addition to this, Suma et al. demonstrated dual tasks could serve as valuable indicators for assessing the impact of VR and cybersickness on working memory, attention, and cognitive load. Although many studies have shown that cybersickness is often reported as lower during real walking tasks[63] compared to stationary tasks, cybersickness can still be present in real walking tasks[35]. Thus, cybersickness during real walking tasks may still affect user experience, task performance, and potentially cognitive load.

As a case study to demonstrate the utility of the dataset, we have evaluated the accuracy of several deep-learning models in classifying cybersickness as most of the currently available datasets are focused on multimodal cybersickness prediction[18]. Additionally, we have conducted SHAP (SHapley Additive exPlanations) explainable-AI analysis to identify the primary features utilized by the deep learning models. In short, our contribution includes:

- VRWalking - A novel open dataset including VR images, eye-tracking, head-tracking, Heart Rate(HR), Galvanic Skin Response(GSR), cybersickness, mental load, physical load, working memory, and attention for participants navigating a maze via real walking inside a VE. See figure 1.
- Data analysis to investigate relationships among cybersickness, mental load, physical load, working memory, and attention for participants navigating a maze via real walking inside a VE.
- Deep learning models trained with VRWalking to classify cybersickness while walking effectively.
- SHAP analysis to identify dominant features for classifying cybersickness while walking.

The deep learning models we applied achieved an impressive 95% accuracy in predicting cybersickness while walking. Surprisingly, our analysis did not identify a correlation between physiological factors and the other metrics. However, the SHAP analysis revealed that physiological data remains some of the dominant features in predicting cybersickness, alongside eye tracking.

The rest of the paper is outlined as follows: 2) Background - a description of the related work in our area and the research gaps present in current datasets, 3) Data collection procedure - the details of how the data was collected and the tasks that participants performed, 4) Data Collected - a description of each type of data that was collected, 5) Data Analysis - descriptive statistics, correlations, and analysis of variance within the data, 6) Discussion of the Data Analysis, 7) Case studies in classification - several deep learning models are evaluated for classifying cybersickness, 8) Classification discussion, 9) Limitations and Future work, and 10) Conclusion.

2 Background

In this section, we define cybersickness, cognitive load, attention, and working memory, discuss measurement and prediction methods, and highlight the research gaps at the intersection of these areas.

2.1 Cybersickness

Cybersickness, a term coined by Stanney et al. [54], encompasses a range of discomforts experienced by users during virtual environment interactions. Rooted in the sensory conflict theory [32], cybersickness arises from inconsistencies between visual and vestibular senses. Studies, such as Mayor et al. [40], have indicated reduced cybersickness severity during real walking locomotion in VR, corroborating this theory. Beyond neuro-physiological perspectives, theories like dual-task interference [46] suggest that concurrent task performance exacerbates discomfort. Kasper et al. [21] noted impaired performance when coordinating tasks in VR. While subjective assessments like the Simulator Sickness Questionnaire (SSQ) [22] and Fast Motion Sickness (FMS) scale [23] gauge cybersickness severity, physiological and motion data offer objective insights [26]. Machine learning approaches, as demonstrated by Oh et al. [44] and Islam et al. [18], leverage both subjective and objective data for cybersickness prediction. However, prior research primarily involved stationary or minimally moving users. Our study, in contrast, focuses on continuous real walking locomotion in VR.

2.2 Cognitive Load

Cognitive Load encompasses the mental and physical effort required to accomplish a task. Cognitive Load Theory (CLT) categorizes cognitive load into subtypes, including intrinsic load or physical load(related to the inherent task difficulty), extraneous load or mental load (involving mental demands), and germane load (related to long-term memory)[57]. Much of the research on working memory has primarily centered on reducing extraneous load, as modifying instructional methods seems like a pragmatic approach to alleviating cognitive requirements[52]. Researchers have previously used the Paas Scale to measure mental load[1, 56] - a single 9-point Likert scale question rating mental load. But as pointed out by Aldekhyl et al., NASA TLX[13] is a better measure to understand the subtypes of the cognitive load as defined by the CLT[1].

2.3 Working Memory & Attention

Working memory is a brain mechanism tasked with temporarily storing and manipulating information [4]. Working memory has limited capacity and can hold only a modest amount of information, whether it be abstract concepts or countable objects[9]. To assess working memory, we employed a methodology akin to the digit span test[6]. The connection between cognitive load theory

and working memory models is somewhat tenuous and may exhibit inconsistencies[53]. Schuler et al. stressed the significance of including working memory as a control variable in research.

Attention is defined as a set of processes that help people comprehend information by giving some elements of the environment (or tasks) priority over others[2]. The study of human visual attention in VR can be done using the data that is readily available in the HMD, such as the eye-tracking, head-tracking data, and the stereoscopic video data[59] or through the use of dual-task metrics[14]. However, little research has been done to explore the relationship between attention and the impact of cybersickness on attention while users are walking. Furthermore, contemporary VR researchers have increasingly emphasized the prediction of multimodal visual attention[61, 11, 34]. However, the availability of labeled data for attention prediction in existing datasets is notably limited.

2.4 VR datasets

See table 1 for an overview of the data types collected in prior open datasets for VR. In the VR domain, datasets typically specialize in specific areas like cybersickness or cognitive load, but not typically both. For instance, while SET[17] and VR.net[64] extensively cover cybersickness, others like VREED[58] and the GW Dataset[27] touch on cognitive load to varying degrees. Dell et al.[10] proposed a machine learning approach to predict cognitive workload. Wan et al.[62] used EEG signals to measure cognitive ability. Hadi et al. and Li et al.[43, 33] assess the dual-task performance and multitasking impact on cognitive status, but they only focused on the older adult population. However, there's a clear gap in datasets that comprehensively capture factors such as physical load, mental load, attention, and working memory, offering a holistic view of cognitive status alongside cybersickness.

To bridge this gap, our VRWalking dataset aims to provide detailed labeling, including complex head and eye tracking data, physiological measures, and left-eye VR images. Additionally, pre and post-session Simulator Sickness Questionnaires (SSQs)[22] are used to assess cybersickness comprehensively. Moreover, a post-session NASA-TLX questionnaire is employed to evaluate cognitive load. To enhance labeling accuracy, Pass-scale-like[49] questions are also administered during the session, enriching the collected time-stamped data.

Importantly, the aforementioned datasets overlook two critical aspects inherent in real-world VR applications: navigation based on actual walking and multitasking. Many contemporary VR applications, such as firefighter response training, medical simulations, and gaming, rely on real walking for navigation and involve multitasking scenarios. To effectively evaluate VR and ensure its accessibility for all users, it's crucial to acknowledge these factors during data collection.

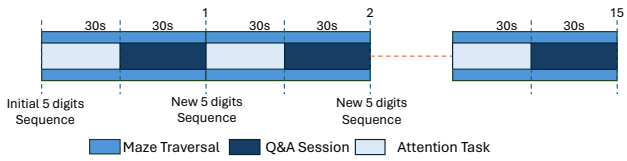

Figure 2: Overview of the Maze Navigation Session

We collected the data for our VRWalking dataset while participants navigated virtual mazes via real walking for 15 minutes while concurrently performing multiple cognitive tasks. The purpose of this data collection was to investigate the relationships between cybersickness, working memory, attention, and cognitive load. Figure 2 shows the timeline of the 15-minute session. In the following sections, we provide details of our hypotheses, the virtual environment, the tasks, the study procedure, metrics, and analysis.

3.1 Participants

We utilized 36 of 39 (23 M, 16 F) participants because the three removed had incomplete sensor data. The participants were recruited from the local area and had a mean age of 25.67 with a standard deviation of 7.22. The participants came from various demographic backgrounds. Twenty-five of 39 participants said they had previously used VR equipment, while thirteen had no prior experience of using VR and one decided not to answer.

3.2 Hypotheses

The primary focus of this data collection was to help determine VR's effects on cybersickness, working memory, physical load, and mental load during multiple cognitive tasks and real walking tasks in a VR maze. Our hypotheses were motivated by effects seen in previous literature, which focused primarily on stationary VR. (see background and related work) H1: Participants will have a significant rise in cybersickness levels proportionate to their time in VR. H2: Working memory will decline, and physical and mental load will increase over time in VR.

3.3 Virtual Environment

The virtual environment consisted of a randomly generated maze, based on a hunt-and-kill algorithm[1] as follows: 1) Build a completely walled maze. 2) Start at the beginning of the previous maze's finish. 3) Hunt in a random direction to find a grid location that has not been visited. 4) Break walls between the platforms until the entire maze has been visited. The maze contained high stone walls, stone platforms, a grey sky, and no ceiling. The maze was a 4x4 grid that randomly spawned the trophy, the finish condition that transports the user to the next randomly generated maze(Figure 1 middle), each with a minimum navigation distance of at least half of the maze. This ensures that the user has to explore at minimum half of the maze in order to progress to the next level. Thus, all randomly generated mazes have approximately the same difficulty. The virtual maze occupies 3m x 3m of real space, with each platform being about .5m x .5m.

3.4 Study Tasks

The primary task was navigating mazes for 15 minutes. See figure 1 left for an example of a maze. As described in figure 2, the whole session was divided into one-minute sections where the maze navigation goes on continuously. In each section, for the first 30 seconds, the participants navigate the maze and complete attention task; in the next 30 seconds, the participants continue navigating the maze, but instead of going through the attention task, they go through a Q&A session. Each maze contained an average of 4-5 turns, and we instructed the user to stay in motion at all times. This task was to be done for the entire duration of 15 minutes unless the participant decided to quit early. We chose 15 minutes as prior literature has shown that most people will have feelings of cybersickness after 10-15 minutes[42].

3.5 Apparatus

We utilized an HTC-Vive Pro Eye headset to present the virtual maze, offering a display resolution of 1440 x 1600 per eye and operating at a refresh rate of 90Hz. The headset provided a wide field of view, spanning 110 degrees. Audio was delivered through the integrated Vive headphones, and to ensure a seamless VR experience, we configured the HMD with the Vive wireless adapter. The

[1] https://tinyurl.com/yvr56h9d

computer used for rendering had the Windows 10 operating system, a 4.35 GHz Intel Core i7 processor, 32 GB DDR3 RAM, and an NVIDIA GeForce RTX 2080 graphics card. The virtual environment was crafted using Unity 3D. To capture eye-tracking data, including gaze direction and pupil diameter, we employed the HTC SRanipal SDK and Tobii HTC Vive Devkit. Additionally, we used the SteamVR lighthouse system to capture head tracking. In order to gather all the data efficiently without interrupting the VR simulation, we used a multi-threaded logging [2] module within the Unity software[31]. We employed the Neulog[3] system to gather heart rate and galvanic skin response data, utilizing two external modular sensors. To ensure convenient data preservation, we employed a Neulog Wi-Fi module for wireless data transmission.

3.6 Procedure

3.6.1 Study Introduction

Participants read the informed consent form, and we gave them an overview of the different activities during the study. They were also told that they would be able to end the study at any moment if they became too uncomfortable. Upon signing the consent form, participants were then instructed to fill out a background questionnaire, which included an initial SSQ, demographic questions, and disability questions. After fitting participants with the HTC Vive Pro Eye headset, physiological monitoring equipment, and controllers, they were brought into a tutorial that explained the tasks throughout the maze.

3.6.2 Tutorial

Before the 15-minute VR session, the participants went through a brief 1-minute virtual tutorial explaining the tasks they had to complete. This tutorial included instructions on operating the handheld controller and which button to press during the attention task. The tutorial also ensured that the participants heard the instructions clearly through the HMD's headphones. In the tutorial, the participants were placed in a flat space, and an audio message explained the questions asked during the maze task. After the introductory message was played, the participant was instructed to press the trigger when they heard a keyword to train them for the attention task. They were then instructed to intersect their hand with a 'trophy' (figure 1 middle) to complete a maze.

3.6.3 Maze Navigation

After this training, the participants were then put through an eye calibration to ensure that the system accurately interprets and records eye movements. The participants were given the target word for the attention metric and the first sequence of five randomly generated digits, as shown in figure 2. Once the participants were ready, they were instructed to start navigating the maze. Every 30 seconds, the Q&A session started, and they answered the three Likert scale questions on cybersickness, mental load, and physical load, recited the previous five digits given, and were given a new five-digit number to remember. The participants navigated the maze via real walking. They also were instructed to complete the attention task throughout the session as they navigated the maze. When at the end of a maze, they intersected the maze completion marker with their controller and then were teleported to the next maze.

3.6.4 Post-experience questionnaires

After fifteen minutes had passed or the participant decided to quit, they exited the maze and completed another SSQ and a NASA TLX questionnaire. At the end of the session, the participant was paid $30 per hour and parking fees.

[2]https://tinyurl.com/yckknkdw

[3]https://neulog.com/

4 Data Collected

We collected the following data for our novel VRWalking dataset:

Table 1 highlights the comprehensive nature of our dataset, encompassing a wide array of features. This dataset provides essential elements for investigating cybersickness, cognitive load, and can be seamlessly integrated into other open datasets to create a more expansive resource for future researchers.

Attention: While inside the VR session, as the participants were navigating through the maze, they heard one word at a time every 2-3 seconds randomly generated from a pool of five words ("Alpha", "Bravo", "Charlie", "Delta", and "Echo"). The participants were given a specific target word at the beginning of the session. Throughout the session, they had to press the instructed button when they heard the words and skip if the word they heard matched the given target word. Two performance metrics were calculated from this task to assess the participant's attention[14].

- Correct Button Presses: We calculated this as a percentile of how many button presses were correct. The correct button presses included presses on other words and skips on target word. We refer to this as *Attention (Success Rate)*.

- Reaction Time: We calculated this as an elapsed time from the word played through the headset and the button presses. This time was then averaged in one-minute windows aligned with the collection of FMS, physical load, and mental load questions. We refer to this as *Attention (Reaction Time)* reported in seconds.

Working memory load: Alongside the target word for the attention task, the participants also received a five-digit number played through the headphones. During the Q&A session, we ask them to repeat the five digits, and at the end of the Q&A session, we give them a new randomly generated five-digit number. The working memory was calculated as a percentile of the number of digits they repeated correctly. The repeated digit was considered correct if the index and the digit matched the given digit. For example, if the given five-digit number is 57342 and the participant responded 58352, we consider this 5X3X2, where the X represents the missed digits. The performance score will become 60%. This task is similar to the digit span test[6] used in previous research[25].

NASA-TLX: We took this after the partipants exited the VR session. It consists of six subscales that assess different aspects of workload: mental demand, physical demand, temporal demand, performance, effort, and frustration. Each subscale is rated on a 0-100 scale, with higher scores indicating a higher workload[13].

Physical Load: We asked a single question from the NASA-TLX once per minute: How physically demanding is the task on a scale of not demanding/1-very demanding/10?

Mental Load: Similar to the Paas scale [57], we asked the following question once per minute: How mentally demanding is the task on a scale of not demanding/1-very demanding/10?

Fast Motion Sickness Scale (FMS): We asked the following question once per minute: How sick do you feel on a scale of no cybersickness/1-very intense cybersickness/10? [23, 12]

SSQ: We took pre-session and post-session SSQ before and after the VR session respectively. From the 16 questions, we calculated the total score and three subscores: Nausea, Oculomotor, and Disorientation[22].

Eye Tracking: The SRanipal SDK(Sensory Reality Affective Neural-integrated Processing SDK) seamlessly incorporates Tobii's eye tracking technology. Through this technology, we gathered detailed eye-tracking data for each participant. Importantly, this eye-tracking data was logged along with timestamps and was automatically stopped when the VR session concluded. The eye-tracking data was sampled at a rate of 60 Hz.

Dataset	Exposure Time	Eye Tracking	Head Tracking	HR	GSR	Locomotion	Cognitive Load	Cybersickness
SET[17]	7 Minutes	Yes	Yes	Yes	Yes	No	No	Yes
VRWalking	**15 Minutes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
VREED[58]	1-3 Minutes	Yes	No	No	Yes	No	No	No
GW Dataset[27]	5 Minutes	Yes	Yes	No	No	Yes	No	No
EHTask[15]	150 Seconds	Yes	Yes	No	No	No	No	No
VR.net[64]	10 Minutes	Yes	Yes	Yes*	No	Mixed	No	Yes

Table 1: Some of the prior open datasets for VR and the data types included. '*' represents partially collected

Head Tracking: For head tracking, we employed the SteamVR Lighthouse tracking system, which consists of base stations emitting laser signals. These signals are picked up by the headset, enabling the system to accurately calculate the headset's position and orientation, thereby providing head-tracking data. This head-tracking data was sampled at a rate of 60 Hz.

Physiological Readings: We gathered bio-physiological data (i.e., Heart Rate(HR) and Galvanic Skin Response (GSR)) using external sensors that were affixed to the participant's fingers. To ensure participants had the freedom to move around comfortably, we employed wireless modular sensors provided by Neulog. This data was collected at a 10 Hz sampling rate.

VR Images: We recorded point-of-view video footage of participants within the VE they were experiencing. These videos were easily accessible through the HMD and were recorded and saved independently.

Given that head tracking and eye tracking data were sampled at a rate of 60 Hz, while HR and GSR were sampled at 10 Hz, we synchronized their frequencies by downsampling all signals to 1 Hz. This synchronization facilitated coherent analysis across modalities and was necessary for creating the dataset for cybersickness prediction.

5 Data Analysis

5.1 Descriptive Statistics

Descriptive statistics can be found in Tables 2 and 3. An example of how these data vary over time can be observed in figure 3. Based on the descriptive statistics for attention, it appears relatively stable with a low standard deviation. However, Figure 3 visually demonstrates that while FMS, physical load, mental load, and Attention (Reaction Time) steadily increase, working memory and Attention (Success Rate) exhibit more fluctuations over time. We also noticed high variability in the physiological measurements in general.

Table 2: Descriptive Statistics

Metric	Mean	SD
FMS	1.93	1.27
HR	78.13	14.25
GSR	3.07	2.4
Physical Load	2.11	1.24
Mental Load	3.75	1.86
Working Memory	76.07	17.98
Attention (Success Rate)	63.91	5.3
Attention (Reaction Time)	0.22	0.13

Furthermore, we can deduce from Table 2 that, on average, the majority of participants experienced a low level of cybersickness as expected for a walking task. Conversely, most participants exhibit elevated levels of mental and physical load compared to cybersickness.

To construct Figure 3 we aggregated the physical load, mental load, and task performance metrics for working memory and attention by computing the average value at each time point across all participants. This means that, for example, the physical load at a given time point represented the average of the physical load values of all 36 participants at that specific time point. Since the value ranges of these metrics differ (e.g., physical load ranges from 1 to 10 while working memory ranges from 0 to 100), we standardized all values to a scale of 0 to 1. This standardization was done to ensure that all metrics could be visually compared on the same scale in the graphs, facilitating a comprehensive overview of the progression of each metric over time (See Figure 3).

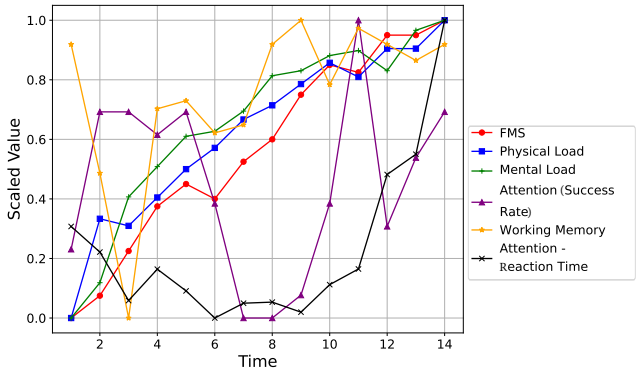

Figure 3: Normalized FMS, Physical Load, Mental Load, Working Memory, Attention (Success Rate), and Attention (Reaction Time) Over Time for one participant

Table 3: Descriptive Statistics of Pre-session and Post-session SSQ

	Pre-session	Post-session
Nausea	0.39±0.69	2.75±1.92
Oculomotor	0.97±1.32	3.94±3.57
Disorientation	0.36±0.68	2.33±2.45
Total Score	6.4±9.32	33.76±27.21

5.2 Differences Over Time

A Shapiro-Wilk test on the data indicated that we cannot assume normality. That is why we performed the Wilcoxon signed-rank test to determine the difference between the pre-session and post-session SSQ subscores and the total scores. All cybersickness subscores indicated a significant increase post-session relative to pre-session: Nausea ($V = 16.0$, $p = 2.02 \times 10^{-6}$, Cohen's $d = 6.16$), Oculomotor ($V = 71.0$, $p = 3.47 \times 10^{-5}$, Cohen's $d = 15.69$), Disorientation ($V=18.0$, $p = 3.41 \times 10^{-5}$, Cohen's $d = 14.89$) and Total Score ($V = 57.0$, $p = 2.60 \times 10^{-7}$, Cohen's $d = 9.12$). The mean scores of each subscore are displayed in Table 3.

We expanded our analysis by stratifying the data based on VR exposure time and participants' FMS scores, conducting Wilcoxon-signed rank tests for paired groups and Mann-Whitney U tests for unpaired groups to assess statistical significance (Table 4). In the case of exposure time, we divided the 15-minute sessions into two

groups of 7.5 minutes each, comparing FMS and physiological data between them. Results revealed statistical significance between the first and second half of FMS scores, with no significant differences observed in heart rate (HR) and galvanic skin response (GSR). However, when examining the initial and final 30 seconds of HR and GSR, significant differences emerged for GSR, suggesting an impact of exposure time on FMS and GSR.

5.3 Differences Between Sick and Non-Sick Participants

Additionally, participants were grouped based on their FMS scores, categorized as 'sick' (FMS score ≥ 2) and 'non-sick' (FMS score < 2), with the latter serving as the lower quartile (Q1) reference. Despite observing no significant differences in physiological data between these groups, significant disparities were detected in NASA-TLX's physical demand and annoyance scales (Table 5).

Table 4: Differences in Data over Time from Wilcoxon Signed Rank Tests

Category	Stats	Value
First Half FMS vs Second Half FMS	W	20
	z	-4.92
	p	0.0009
	Cohen's d	18.33333333
First Half Physical Load vs Second Half Physical Load	W	23
	z	-4.87
	p	0.0001
	Cohen's d	13.41
First Half Mental Load vs Second Half Mental Load	W	17.5
	z	-4.96
	p	3.64×10^{-6}
	Cohen's d	5.75
First 30s GSR vs Second 30s GSR	W	93
	z	-3.77
	p	6.62×10^{-5}
	Cohen's d	15.5

Table 5: Mann-Whitney U tests comparing NASA-TLX (physical demand, annoyance), physical load, mental load, Working Memory, and attention between sick participants (13) and non-sick participants(23)

Category	Mean	SD	z-value	U	p-value	Cohen's d
Physical Demand-NASA-TLX (Sick)	3.15	1.63	2.50	225.5	0.009	0.43
Physical Demand-NASA-TLX (Non-sick)	1.87	1.01				
Annoyance-NASA-TLX (Sick)	5.07	2.22	3.80	265	0.0001	0.50
Annoyance-NASA-TLX (Non-sick)	2.26	1.05				
Working Memory (Sick)	65.27	22.68	-2.10	85.5	0.03	0.16
Working Memory (Non-sick)	82.17	11.22				

6 Data Analysis Discussion

The data analysis section primarily aimed to validate prior hypotheses established across various data collection sessions, thereby demonstrating the comparable potential and applicability of the collected data. We primarily conducted correlation tests to determine the relationship between physical load and mental load encountered during VR sessions, revealing a positive monotonic correlation between them (Coefficient = 0.63). Furthermore, the FMS was validated against post-session SSQ subscores, revealing a positive correlation between FMS and all the subscores. The significance tests were employed to identify features significantly associated with the FMS or other subjective measures gathered during the session. This process provided valuable insights into the features that should be prioritized when collecting similar data in the future. This same motivation drove the subsequent SHAP analysis, which further highlighted the impact of individual features on the target outcomes.

6.1 Cybersickness

We have observed lower post-session subscores compared to those reported in a previous study[18]. Our hypothesis is that this reduction may be attributed to the continuous movement experienced during the VR session, which could potentially counteract the sensory conflict that often leads to cybersickness. Despite the reduction in cybersickness associated with real walking-based navigation, it's noteworthy that 13 participants (36%) exhibited an average FMS higher than 2, with 21 participants (58%) showing an average FMS higher than 1.14 (median).

We notice a notable variation in FMS across time, as evidenced by the significance test conducted on the FMS values for the first and second halves. This outcome aligns with the findings from the SSQ assessments conducted before and after the session, where we identified substantial disparities in all subscores as well as the total scores. This collectively suggests that prolonged exposure to virtual reality leads to the onset of cybersickness.

We conducted additional analyses to explore the relationships between cybersickness and the collected features. Notably, we identified a robust correlation between the post-session SSQ subscores and the participants' average FMS, affirming the utility of FMS for a detailed understanding of cybersickness. Conversely, we did not uncover significant correlations between HR, GSR and FMS.

6.2 Mental Load and Physical Load

Both physical load and mental load showed significant variations when the two groups were categorized based on time. Furthermore, when evaluating the reported physical demand using the NASA TLX, there was a discernible contrast between the non-sick (23 participants) and sick groups (13 participants), with a substantial increase in mean values observed in the latter. However, the verbal reports of physical load and mental load did not exhibit significant variations between the non-sick and sick groups.

We had anticipated differences in both physical load and mental load between the sick and non-sick groups. As expected, we observed the anticipated outcome in the reported physical demand from the NASA-TLX. Both FMS and physical load, as well as mental load, exhibited significant variations over time. However, the absence of differences in physical load and mental load between the sick and non-sick groups suggests that physical and mental load may not be closely related to cybersickness.

6.3 Working Memory and Attention

Table 5 highlights a statistically significant difference in working memory between the sick and non-sick groups. However, no significant difference was observed for attention metrics. In Figure 3, an upward trend in reaction time over time suggests reduced attention over time. Furthermore, However, we found no indicate no significant relationship between working memory and time. Additionally, it's worth noting a significant correlation between Attention (Success Rate) and Attention (Reaction Time) (p-value < 0.05) was observed during Spearman correlation analysis, with a coefficient of -0.51.

7 Case Study in Cybersickness Classification

To demonstrate potential use cases for our VRWalking dataset, we performed a case study using deep learning models to classify cybersickness.

7.1 Categorization

To support classification, we categorize the FMS values into three categories based on quantiles.

$$Class_t = \begin{cases} \text{Low}, & \text{if } 0 < X \leq Q_1 \\ \text{Medium}, & \text{if } Q_1 < X \leq Q_2 \\ \text{High}, & \text{if } X > Q_2 \end{cases} \quad (1)$$

In Equation 1, X represents FMS. The quantiles were set as $Q_1 = 2, Q_2 = 3$. The deliberate selection of Q1=2 and Q2=3 was made to enhance the classification of cybersickness effectively. Despite the FMS score range spanning from 1 to 10, opting for a smaller range was informed by prior research[18, 17] and aimed at achieving a more balanced dataset. Typically, FMS levels exceeding 6 are infrequent. Hence, choosing Q1=2 and Q2=3 was ideal for creating a well-balanced classification dataset.

7.2 Neural Network Models: LSTM, GRU & MLP

There are three different neural network architectures we used for the cybersickness classification: LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), and MLP (Multilayer Perceptron). For sequential data, such as time series data, recurrent neural network (RNN) architectures LSTM and GRU were principally developed. MLP is a feedforward neural network design made up of several interconnected layers of neurons that are also capable of handling time-series data. As illustrated in Table 6, the LSTM model comprises two LSTM layers for capturing sequential patterns, each followed by a dropout layer for regularization against overfitting. The final dense layer serves multi-class classification. The GRU model incorporates a GRU layer for capturing temporal dependencies, followed by a dropout layer for regularization. It also features an LSTM layer for complex temporal modeling, another dropout layer, and a second GRU layer with an additional dropout. The final layer is a dense layer for classification. The MLP model starts with an initial dense layer for feature extraction, followed by dropout for regularization. It includes more hidden dense layers with a dropout layer between them. The final layer is a dense layer for multi-class classification. Our first preference for classification mostly focuses on the use of simple models. This tactical choice not only makes it easier to conduct SHAP Analysis but also makes it easier to interpret. It is important to note that applying SHAP analysis to complicated models like the deep fusion models used for a regression task can be extremely difficult and frequently makes it difficult to identify the dominant features[29, 30]. Table 6 illustrates the model architecture we used for the severity classification.

Table 6: Model Architectures for Classification

Architecture for the LSTM Model					
Layer	Type	Output Shape	#Param	Dropout	Activation
1	LSTM	256	264192	0.2	-
2	Dropout	256	0	0.2	-
3	LSTM	128	197120	0.2	-
4	Dropout	128	0	0.2	-
5	Dense	3	387	-	Softmax
Total No. of Params: 461,699					
Architecture for the GRU Model					
1	GRU	32	3360	0.2	-
2	Dropout	32	0	0.2	-
3	LSTM	64	18816	0.2	-
4	Dropout	64	0	0.2	-
5	GRU	128	74496	0.2	-
6	Dropout	128	0	0.2	-
7	Dense	3	387	-	Softmax
Total No. of Params: 97,059					
Architecture for the MLP Model					
1	Dense	32	2368	0.25	ReLU
2	Dropout	32	0	0.25	-
3	Dense	128	4224	0.25	ReLU
4	Dropout	128	0	0.25	-
5	Dense	32	4128	0.25	ReLU
6	Dropout	32	0	0.25	-
7	Dense	8	264	0.25	ReLU
8	Dropout	8	0	0.25	-
9	Dense	3	27	-	Softmax
Total No. of Params: 11,011					

7.3 Training Setup

We employed a 10-fold cross-validation technique to train and assess the performance of our proposed model, following a methodology similar to previous studies [5, 38, 45]. In the k-fold cross-validation process, the dataset is divided into k subsets. One of these subsets is used for testing the model while the remaining (k-1) subsets serve as the training data [19]. This procedure is repeated k times, each time using a different subset as the test data and the rest as the training data. This approach helps mitigate any potential dataset bias [5].

In our data preprocessing approach, we applied data normalization to the eye tracking data, ensuring consistent scaling across all observations. Additionally, we employed exponential smoothing techniques to denoise the HR and GSR signals, drawing inspiration from prior research methodologies focused on cybersickness prediction[17, 18, 16].

To optimize the model parameters during training, we allocated 30% of the training dataset as validation data for each fold [48]. We employed the Adam optimizer with a training duration of 200 epochs and a batch size of 256, all configured through hyperparameter tuning. We used categorical cross-entropy as the loss function. To safeguard against overfitting, we implemented an early-stopping mechanism with a patience setting of 30 during the model training [7].

7.4 Classification Results

The mean accuracy, precision, and recall for 10-fold cross-validation are listed in Table 7. During the classification task, GRU and LSTM models achieved the highest accuracy. For the mental load classification, LSTM outperformed the GRU model, achieving an accuracy of 95%.

7.5 SHAP Analysis

In order to further analyze and better understand the performance of the models, we employed SHapley Additive Explanations (SHAP). SHAP is a post-hoc explanation method that ranks the input features. The goal of SHAP is to explain the prediction of a given DL model (e.g., LSTM, GRU, MLP, etc.) for a given set of input samples (e.g., eye and head tracking, HR, etc.). The model architecture for the used DL models is summarized in Table 6. Mangalathu et al. emphasize that while complex deep learning models can achieve high accuracy, interpreting them through SHAP analysis poses challenges[39]. These models often possess intricate internal representations and intricate feature interactions, rendering the provision of clear and easily understandable explanations for model predictions using SHAP a daunting task. That is why we have used some simpler DL models in our classification, and the global explanation done by the SHAP analysis shows us the dominant features for cybersickness classification. We have listed the DL models that achieved the best performance during the classification task in table 7. Although during the correlation test, we did not find any significant correlation between HR, GSR, and FMS, from

Table 7: Mean Accuracy, Precision and Recall of the 10-Fold Cross Validation on Cybersickness Classification for Different Models

Type	Used Models	% Accuracy	% Precision			% Recall		
			Low	Medium	High	Low	Medium	High
CS_t	GRU	95	97	92	95	98	88	93
	LSTM	95	96	90	97	98	89	90
	MLP	82	98	62	69	83	72	97

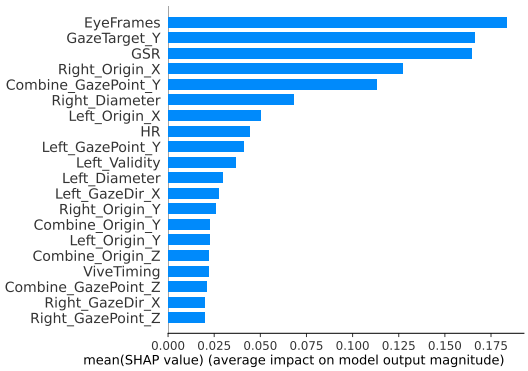

Figure 4: Cybersickness Classification (GRU)

the SHAP analysis, we can see that both HR and GSR were dominant features when predicting Cybersickness Figure 4 illustrates the dominant features.

8 Cybersickness Classification Discussion

The LSTM, GRU, and MLP models demonstrated outstanding predictive performance for all attributes, achieving exceptional accuracy. Notably, the MLP had a lower performance compared to the LSTM and GRU models, but it's important to highlight that the GRU had 8.8 times more parameters, and the LSTM had 42 times more parameters than the MLP. This led to significantly reduced training times for the MLP.

When employing the basic LSTM, GRU, and MLP models, we observed a cybersickness severity classification on par with state-of-the-art results, such as Kundu et al.'s [30] achievement of 94% accuracy with an LSTM. Despite the data collection occurring in a novel environment with navigation and cognitive tasks and the absence of initial resting baseline measurements for data normalization, our deep learning models trained with our dataset delivered an outstanding performance, achieving an impressive accuracy rate of 95%. The LSTM and GRU models we used exhibited state-of-the-art level precision and recall for the individual classes compared to prior research[30]. We anticipate that incorporating a resting baseline and employing more sophisticated models could yield even better outcomes, especially considering that previous research demonstrated the superiority of deep fusion models over basic deep learning models[16]. We initially anticipated that the swift movements associated with the navigation task might disrupt the collection of physiological data. Surprisingly, after employing exponential smoothing to standardize the readings, they emerged as some of the most influential features in predicting cybersickness. Additionally, we speculated that the navigational task compelled participants to engage in frequent head movements, introducing considerable variability in the head-tracking data. This may explain why head tracking was not as effective in predicting cybersickness as observed in prior research[17, 28].

9 Limitations and Future Work

Our dataset was not balanced with respect to gender, which is known to correlate with cybersickness[37, 41] and potentially other data types that we collected. To address this, we intend to collect data from a more diverse and gender-balanced participant group to improve generalizability. Furthermore, the attention metric we utilized is limited in that it does not provide insights into sustaining visual attention[50]. Therefore, we aim to incorporate metrics that assess reaction time from eye tracking to gain a better understanding of sustaining visual attention.

Additionally, we exclusively utilized basic deep learning models in this study. However, we anticipate achieving improved results by employing advanced models, such as deep fusion models, which may be better equipped to learn from time series data[16]. Moreover, we only used a maze virtual environment with real walking in the data collection. We plan to evaluate different virtual environments with different locomotion interfaces in later studies.

In the future, our aim is to extend the use of these models to classify cognitive effects on working memory and attention and for precise value prediction through regression. Moreover, in line with previous research, we plan to explore the application of SHAP for model reduction. This would enable us to develop efficient and lightweight models suitable for deployment on VR headsets. This approach would empower us to effectively predict both cybersickness and cognitive load without the need for external hardware. Moreover, we plan on doing regression instead of classification as that is better suited for cybersickness prediction and and for the classification task we plan on conducting statistical analysis to calculate the categorization boundary for a better generalizable approach.

10 Conclusion

In conclusion, we collected a dataset to understand the relationship between cybersickness, attention, mental load, physical load, and working memory in VR. Participants were tasked with maze navigation via real walking for 15 minutes. Throughout the experiment, we collected head-tracking, eye-tracking, HR, GSR, and self-reported data on cybersickness, mental load, and physical load while users performed working memory and attention tasks. Results suggested that participants grew more sick over time. Moreover, the participants who were sicker experienced a significantly higher demand on their working memory performance and their self-reported physical load than participants who were less sick. Next, as an example of the use of the dataset, we developed several deep learning models with the intention of predicting cybersickness based on head tracking, eye tracking, GSR, and HR sensor data as input. Using several simple deep learning models, GRU, LSTM, and MLP, we were able to effectively classify cybersickness with 95% accuracy. Lastly, we performed a SHAP analysis to identify which features impacted our classifiers the most. Ultimately, VR developers could utilize our dataset to create virtual environments and predictive models that can effectively adapt to the user's current state of mind, which could lead to improved and more personalized VR applications.

Acknowledgments

This work was supported in part by Army Research Lab awards W911NF2310401 and W911NF2420035, and National Science Foundation awards 2211785 and 2316240.

REFERENCES

[1] S. Aldekhyl, R. B. Cavalcanti, and L. M. Naismith. Cognitive load predicts point-of-care ultrasound simulator performance. *Perspectives on medical education*, 7:23–32, 2018. 2

[2] A. Allport. Attention and control: Have we been asking the wrong questions? a critical review of twenty-five years. 1993. 3

[3] A. Armougum, E. Orriols, A. Gaston-Bellegarde, C. Joie-La Marle, and P. Piolino. Virtual reality: A new method to investigate cognitive load during navigation. *Journal of Environmental Psychology*, 65:101338, 2019. 2

[4] M. Baddeley. Herding, social influence and economic decision-making: socio-psychological and neuroscientific analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1538):281–290, 2010. 2

[5] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Advances in Neural Information Processing Systems*, 16, 2003. 7

[6] H. L. Blackburn and A. L. Benton. Revised administration and scoring of the digit span test. *Journal of consulting psychology*, 21(2):139, 1957. 2, 4

[7] R. Caruana, S. Lawrence, and C. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in neural information processing systems*, 13, 2000. 7

[8] D. Checa and A. Bustillo. A review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools and Applications*, 79(9):5501–5527, 2020. 2

[9] N. Cowan. Working memory underpins cognitive development, learning, and education. *Educational psychology review*, 26:197–223, 2014. 2

[10] F. Dell'Agnola, N. Momeni, A. Arza, and D. Atienza. Cognitive workload monitoring in virtual reality based rescue missions with drones. In *International conference on human-computer interaction*, pp. 397–409. Springer, 2020. 3

[11] F. Fathy, Y. Mansour, H. Sabry, M. Refat, and A. Wagdy. Virtual reality and machine learning for predicting visual attention in a daylit exhibition space: A proof of concept. *Ain Shams Engineering Journal*, 14(6):102098, 2023. 3

[12] J. P. Freiwald, Y. Göbel, F. Mostajeran, and F. Steinicke. The cybersickness susceptibility questionnaire: predicting virtual reality tolerance. In *Proceedings of Mensch Und Computer 2020*, pp. 115–118. 2020. 4

[13] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988. 2, 4

[14] C. Ho and C. Spence. Assessing the effectiveness of various auditory cues in capturing a driver's visual attention. *Journal of experimental psychology: Applied*, 11(3):157, 2005. 3, 4

[15] Z. Hu, A. Bulling, S. Li, and G. Wang. Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(4):1992–2004, 2023. doi: 10.1109/TVCG.2021.3138902 5

[16] R. Islam, K. Desai, and J. Quarles. Cybersickness prediction from integrated hmd's sensors: A multimodal deep fusion approach using eye-tracking and head-tracking data. In *2021 IEEE international symposium on mixed and augmented reality (ISMAR)*, pp. 31–40. IEEE, 2021. 2, 7, 8

[17] R. Islam, K. Desai, and J. Quarles. Towards forecasting the onset of cybersickness by fusing physiological, head-tracking and eye-tracking with multimodal deep fusion network. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 121–130. IEEE, 2022. 3, 5, 7, 8

[18] R. Islam, Y. Lee, M. Jaloli, I. Muhammad, D. Zhu, and J. Quarles. Automatic detection of cybersickness from physiological signal in a virtual roller coaster simulation. In *2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*, pp. 648–649. IEEE, 2020. 2, 6, 7

[19] G. James, D. Witten, T. Hastie, R. Tibshirani, et al. *An introduction to statistical learning*, vol. 112. Springer, 2013. 7

[20] M. Javaid and A. Haleem. Virtual reality applications toward medical field. *Clinical Epidemiology and Global Health*, 8(2):600–605, 2020. 2

[21] R. W. Kasper, H. Cecotti, J. Touryan, M. P. Eckstein, and B. Giesbrecht. Isolating the neural mechanisms of interference during continuous multisensory dual-task performance. *Journal of cognitive neuroscience*, 26(3):476–489, 2014. 2

[22] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993. 2, 3, 4

[23] B. Keshavarz and H. Hecht. Validating an efficient method to quantify motion sickness. *Human factors*, 53(4):415–426, 2011. 2, 4

[24] A. Kim, J.-E. Lee, and K.-M. Lee. Exploring the relative effects of body position and locomotion method on presence and cybersickness when navigating a virtual environment. *ACM Transactions on Applied Perception*, 21(1):1–25, 2023. 2

[25] B. R. Kim, M. H. Chun, L. S. Kim, and J. Y. Park. Effect of virtual reality on cognition in stroke patients. *Annals of rehabilitation medicine*, 35(4):450–459, 2011. 4

[26] Y. Y. Kim, H. J. Kim, E. N. Kim, H. D. Ko, and H. T. Kim. Characteristic changes in the physiological components of cybersickness. *Psychophysiology*, 42(5):616–625, 2005. 2

[27] R. Kothari, Z. Yang, C. Kanan, R. Bailey, J. B. Pelz, and G. J. Diaz. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports*, 10(1):2539, 2020. 3, 5

[28] R. K. Kundu, O. Y. Elsaid, P. Calyam, and K. A. Hoque. Vr-lens: Super learning-based cybersickness detection and explainable ai-guided deployment in virtual reality. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 819–834, 2023. 8

[29] R. K. Kundu, R. Islam, P. Calyam, and K. A. Hoque. Truvr: Trustworthy cybersickness detection using explainable machine learning. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 777–786. IEEE, 2022. 7

[30] R. K. Kundu, R. Islam, J. Quarles, and K. A. Hoque. Litevr: Interpretable and lightweight cybersickness detection using explainable ai. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 609–619. IEEE, 2023. 7, 8

[31] M. Lamb, M. Brundin, E. Perez Luque, and E. Billing. Eye-tracking beyond peripersonal space in virtual reality: validation and best practices. *Frontiers in Virtual Reality*, 3:864653, 2022. 4

[32] J. J. LaViola Jr. A discussion of cybersickness in virtual environments. *ACM Sigchi Bulletin*, 32(1):47–56, 2000. 2

[33] X. Li, K. S. Niksirat, S. Chen, D. Weng, S. Sarcar, and X. Ren. The impact of a multitasking-based virtual reality motion video game on the cognitive and physical abilities of older adults. *Sustainability*, 12(21):9106, 2020. 3

[34] X. Li, Y. Shan, W. Chen, Y. Wu, P. Hansen, and S. Perrault. Predicting user visual attention in virtual reality with a deep learning model. *Virtual Reality*, 25(4):1123–1136, 2021. 3

[35] J. Lohman and L. Turchet. Evaluating cybersickness of walking on an omnidirectional treadmill in virtual reality. *IEEE Transactions on Human-Machine Systems*, 52(4):613–623, 2022. 2

[36] T. Luong, A. Pléchata, M. Möbus, M. Atchapero, R. Böhm, G. Makransky, and C. Holz. Demographic and behavioral correlates of cybersickness: A large lab-in-the-field study of 837 participants. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 307–316. IEEE, 2022. 2

[37] C. MacArthur, A. Grinberg, D. Harley, and M. Hancock. You're making me sick: A systematic review of how virtual reality research considers gender & cybersickness. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–15, 2021. 8

[38] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR 2011*, pp. 3177–3184. IEEE, 2011. 7

[39] S. Mangalathu, S.-H. Hwang, and J.-S. Jeon. Failure mode and effects analysis of rc members based on machine-learning-based shapley additive explanations (shap) approach. *Engineering Structures*, 219:110927, 2020. 7

[40] J. Mayor, L. Raya, and A. Sanchez. A comparative study of virtual reality methods of interaction and locomotion based on presence, cy-

bersickness, and usability. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1542–1553, 2019. 2

[41] M. Melo, G. Gonçalves, D. Narciso, and M. Bessa. Impact of different role types and gender on presence and cybersickness in immersive virtual reality setups. In *2021 international conference on graphics and interaction (ICGI)*, pp. 1–8. IEEE, 2021. 8

[42] M. Melo, J. Vasconcelos-Raposo, and M. Bessa. Presence and cybersickness in immersive content: Effects of content type, exposure time and gender. *Computers & Graphics*, 71:159–165, 2018. 3

[43] H. Nobari, S. Rezaei, M. Sheikh, J. P. Fuentes-García, and J. Pérez-Gómez. Effect of virtual reality exercises on the cognitive status and dual motor task performance of the aging population. *International Journal of Environmental Research and Public Health*, 18(15):8005, 2021. 3

[44] S. Oh and D.-K. Kim. Machine–deep–ensemble learning model for classifying cybersickness caused by virtual reality immersion. *Cyberpsychology, Behavior, and Social Networking*, 24(11):729–736, 2021. 2

[45] N. Padmanaban, T. Ruban, V. Sitzmann, A. M. Norcia, and G. Wetzstein. Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos. *IEEE transactions on visualization and computer graphics*, 24(4):1594–1603, 2018. 7

[46] H. Pashler. Dual-task interference in simple tasks: data and theory. *Psychological bulletin*, 116(2):220, 1994. 2

[47] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & education*, 147:103778, 2020. 2

[48] B. D. Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007. 7

[49] J. Roelofs, L. McCracken, M. L. Peters, G. Crombez, G. van Breukelen, and J. W. Vlaeyen. Psychometric evaluation of the pain anxiety symptoms scale (pass) in chronic pain patients. *Journal of behavioral medicine*, 27:167–183, 2004. 3

[50] M. Rosenberg, S. Noonan, J. DeGutis, and M. Esterman. Sustaining visual attention in the face of distraction: a novel gradual-onset continuous performance task. *Attention, Perception, & Psychophysics*, 75:426–439, 2013. 8

[51] E. Sayyad, M. Sra, and T. Höllerer. Walking and teleportation in wide-area virtual reality experiences. In *2020 IEEE international symposium on mixed and augmented reality (ISMAR)*, pp. 608–617. IEEE, 2020. 2

[52] C. Schrader and T. J. Bastiaens. The influence of virtual presence: Effects on experienced cognitive load and learning outcomes in educational computer games. *Computers in Human Behavior*, 28(2):648–658, 2012. 2

[53] A. Schüler, K. Scheiter, and E. van Genuchten. The role of working memory in multimedia instruction: Is working memory working during learning from text and pictures? *Educational Psychology Review*, 23:389–411, 2011. 3

[54] K. M. Stanney, R. S. Kennedy, and J. M. Drexler. Cybersickness is not simulator sickness. In *Proceedings of the Human Factors and Ergonomics Society annual meeting*, vol. 41, pp. 1138–1142. SAGE Publications Sage CA: Los Angeles, CA, 1997. 2

[55] E. Suma, S. Finkelstein, M. Reid, S. Babu, A. Ulinski, and L. F. Hodges. Evaluation of the cognitive effects of travel technique in complex real and virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 16(4):690–702, 2009. 2

[56] J. Sweller. Measuring cognitive load. *Perspectives on medical education*, 7:1–2, 2018. 2

[57] J. Sweller, J. J. van Merriënboer, and F. Paas. Cognitive architecture and instructional design: 20 years later. *Educational psychology review*, 31:261–292, 2019. 2, 4

[58] L. Tabbaa, R. Searle, S. M. Bafti, M. M. Hossain, J. Intarasisrisawat, M. Glancy, and C. S. Ang. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(4):1–20, 2021. 3, 5

[59] E. Upenik and T. Ebrahimi. A simple method to obtain visual attention data in head mounted virtual reality. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 73–78. IEEE, 2017. 3

[60] S. Varmaghani, Z. Abbasi, S. Weech, and J. Rasti. Spatial and attentional aftereffects of virtual reality and relations to cybersickness. *Virtual Reality*, 26(2):659–668, 2022. 2

[61] A. Voinescu, K. Petrini, D. Stanton Fraser, R.-A. Lazarovicz, I. Papavă, L. A. Fodor, and D. David. The effectiveness of a virtual reality attention task to predict depression and anxiety in comparison with current clinical measures. *Virtual reality*, 27(1):119–140, 2023. 3

[62] B. Wan, Q. Wang, K. Su, C. Dong, W. Song, and M. Pang. Measuring the impacts of virtual reality games on cognitive ability using eeg signals and game performance data. *IEEE Access*, 9:18326–18344, 2021. 3

[63] Y. Wang, J.-R. Chardonnet, and F. Merienne. Development of a speed protector to optimize user experience in 3d virtual environments. *International Journal of Human-Computer Studies*, 147:102578, 2021. 2

[64] E. Wen, C. Gupta, P. Sasikumar, M. Billinghurst, J. Wilmott, E. Skow, A. Dey, and S. Nanayakkara. Vr. net: A real-world dataset for virtual reality motion sickness research. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3, 5

[65] D. Žagar, M. Svetina, A. Košir, and F. Dimc. Human factor in navigation: overview of cognitive load measurement during simulated navigational tasks. *Journal of Marine Science and Engineering*, 8(10):775, 2020. 2