

Uncovering EEG Patterns Consistently Associated with Cybersickness Discomfort Using Deep Learning Interpretability Maps

Jacqueline Yau

jhyau2@illinois.edu

University of Illinois Urbana-Champaign

Evan G. Center

University of Oulu

Paweł Augustynowicz

John Paul II Catholic University of Lublin

Steven M. LaValle

University of Oulu

Timo Ojala

University of Oulu

Wenzhen Yuan

University of Illinois Urbana-Champaign

Nancy M. Amato

University of Illinois Urbana-Champaign

Paweł Stróžak

John Paul II Catholic University of Lublin

Minje Kim

University of Illinois Urbana-Champaign

Kara D. Federmeier

University of Illinois Urbana-Champaign

Katherine J. Mimnaugh

University of Illinois Urbana-Champaign

Article

Keywords: Cybersickness, Machine Learning, EEG, P3, P300

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11176708/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Uncovering EEG Patterns Consistently Associated with Cybersickness Discomfort Using Deep Learning Interpretability Maps

Jacqueline Yau^{1*}, Evan G. Center², Paweł Augustynowicz³,
Steven M. LaValle^{1,2}, Timo Ojala², Wenzhen Yuan¹,
Nancy M. Amato^{1,5}, Paweł Strózak³, Minje Kim¹,
Kara D. Federmeier^{4,5}, Katherine J. Mimnaugh^{1*}

^{1*}Siebel School of Computing and Data Science, University of Illinois
Urbana-Champaign, Urbana, IL, USA.

²Faculty of Information Technology and Electrical Engineering,
University of Oulu, Oulu, Finland.

³Department of Experimental Psychology, Institute of Psychology, The
John Paul II Catholic University of Lublin, Lublin, Poland.

⁴Department of Psychology, University of Illinois Urbana-Champaign,
Champaign, IL, USA.

⁵Beckman Institute for Advanced Science and Technology, University of
Illinois Urbana-Champaign, Urbana, IL, USA.

*Corresponding author(s). E-mail(s): jhyau2@illinois.edu;
kmimnau2@illinois.edu;

Contributing authors: evan.center@oulu.fi; pawel.augustynowicz@kul.pl;
steven.lavalle@oulu.fi; timo.ojala@oulu.fi; yuanwz@illinois.edu;
namato@illinois.edu; pawel.strozak@kul.pl; minje@illinois.edu;
kfederme@illinois.edu;

Abstract

Uncomfortable sensations similar to motion sickness, called cybersickness, can develop when using Virtual Reality (VR) head-mounted displays. Cybersickness poses a hindrance to greater use of VR technology. Brain activity recorded using electroencephalogram (EEG) can be used to unintrusively detect cybersickness and the discomfort it causes. To intervene for mitigation, machine learning algorithms that can extract meaningful signals related to cybersickness discomfort from the

rest of the brain data will be required. In this work, we determined which features in EEG data were most helpful for classifying cybersickness-related discomfort by building a framework with neural networks and interpretability maps. Using brain data from two separate auditory event-related potential (ERP) cybersickness user studies, we extracted which spatio-temporal EEG features (from sensor locations and time steps) were most important for discomfort classification. For the first dataset (n=29), across 120 runs of our framework with three different neural networks over multiple random seeds, the models consistently pointed to scalp locations in left frontal, midline central, and right parietal areas as most helpful in determining if EEG data belonged to someone who was experiencing cybersickness discomfort. Similar locations were also tagged in a second dataset (n=36), showing cross-dataset generalizability of our findings across time, VR stimulus, and participant sample. Early periods in the extracted ERP time windows, approximately between 80 and 260 milliseconds post-stimulus, were marked as important for model classification in both datasets. These results help clarify which tagged features can be used for cybersickness-related (and potentially other types of) discomfort classification with EEG in the future. We will provide our code [code released on acceptance].

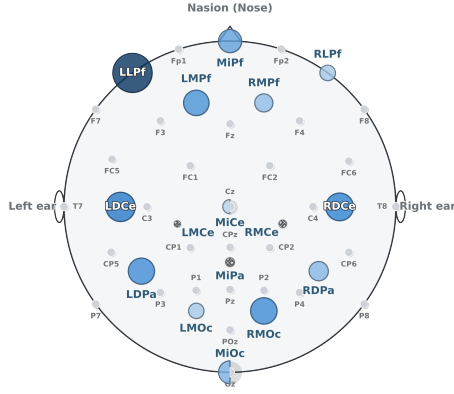
Keywords: Cybersickness, Machine Learning, EEG, P3, P300

1 Introduction

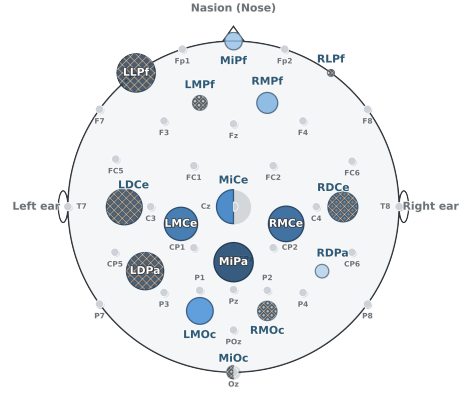
Virtual reality (VR) is an effective medium for training surgeons, educating students, treating pain, and improving mental health [1]. However, the unpleasant symptoms that develop from using VR devices, called cybersickness, preclude long duration applications and makes it difficult for some individuals to use VR at all [2]. At the moment, there are few options for precise passive cybersickness detection. Thus, there has been substantial effort in the scientific community to find methods to unintrusively detect cybersickness using brain activity recorded with electroencephalogram (EEG) [3].

Given the complexity of brain data, it is desirable to find features indicative of cybersickness that can be used by machine learning models (ML) to improve classification. If we can determine which features to focus the models on, it is possible that cybersickness and/or the discomfort that it causes could be detected as it occurs. It might even be possible to detect discomfort from brain activity even before VR users report it, because pain-related sensory processing can precede conscious awareness [4, 5]. However, diagnostic patterns in brain activity could occur at different time points, electrode locations, band power changes, or some combination of the three. Determining which of these features is predictive of discomfort requires overcoming considerable hurdles, including limited datasets for training models, data with high inter-subject variability, and confounds in the signal from unrelated brain activity.

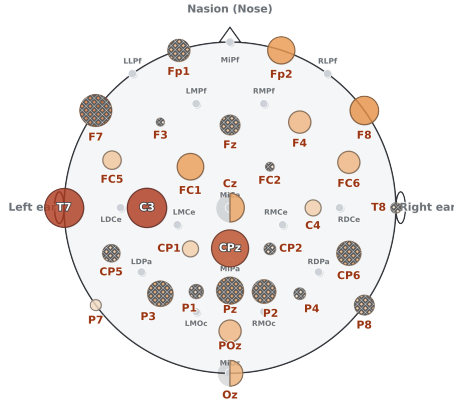
In this article, we describe an extensive investigation of EEG data using a number of supervised deep learning methods, and present the resulting EEG features consistently found to be associated with cybersickness-related discomfort. In overcoming the aforementioned obstacles, we are able to offer several important



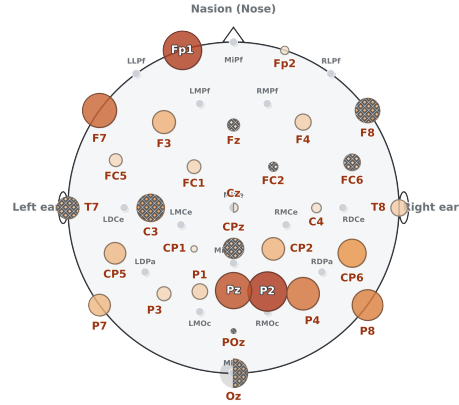
(a) Discomfort Group (ERPVR)



(b) No Discomfort Group (ERPVR)



(c) High Discomfort Group (VectionVR)



(d) Low Discomfort Group (VectionVR)

Fig. 1: Importance of electrode features for discomfort classification across all runs. ERPVR (top) used a 16-channel approximately equidistant montage, and VectionVR (bottom) used a modified 30-channel 10-20 layout. Circle sizes and colors are based on calculated integrated gradients. Circles with crosshatch indicate that feature was negatively contributing to model classification at that magnitude. The top three most important electrodes have labels inside their circles. These integrated gradients are the same as in Figures 3a and 3b.

contributions. First, we provide a feature extraction framework (and its code) that highlights which input features from brain data provide the highest attribution for model classification. While much previous ML work has investigated changes in power in frequency bands like alpha and beta (see [3, 6, 7] for review), here we have built an explainable classification method for EEG data from event-related potential (ERP) studies, a much less studied area with better comparison across samples. Second, although the attribution maps displayed complex interactions across electrodes and time, multiple models across multiple random seeds showed consistent results pointing to the same subset of electrodes as being the most helpful in classifying cybersickness-related discomfort in one dataset (called ERPVR): LLPf, LDCe, RDCe, RMOc, and LDPa for samples with discomfort and MiPa, RMCe, LMCE, MiCe and LMOc for samples without discomfort. We also found related patterns in the second dataset (VectionVR): C3, T7, CPz, F8, and Fp2 for high levels of discomfort and P2, Fp1, Pz, F7, and P4 for low levels of discomfort (see Figure 1). The datasets were collected at different times, in different countries, with different VR headsets, different EEG systems, different electrode locations, and different stimuli in virtual reality. Thus, the results presented here provide a contribution to the scientific literature on important features in auditory ERP data that could be used as tagged features for detection of cybersickness discomfort.

2 Background

2.1 Electroencephalography and Event-Related Potentials

EEG data recorded at the scalp reflect voltage fluctuations generated primarily by summed postsynaptic activity of populations of synchronously active neurons [8]. These signals reflect sensory, motor, cognitive, and other ongoing neural processes. Because EEG reflects many simultaneous processes, activity unrelated to the process of interest contributes “noise” relative to the signal being studied, even when biological, equipment, and environmental artifacts have been well controlled. One method to address this issue of noise involves repeatedly invoking patterns of brain activity in response to a stimulus, eliciting what are known as Event-Related Potentials (ERPs). ERPs can be extracted by averaging across those repetitions to cancel out inconsistent activation [8].

During VR sickness studies, continuous EEG reflects multiple concurrent processes, including processing of the visual environment, optic flow, and activity associated with discomfort. Optic flow can induce *vection*, illusory sensations of self-motion, which is itself an important contributor to cybersickness [9]. Auditory stimuli presented during visual VR experiences provide repeated, precisely time-locked events against this background of ongoing activity. As presentation of the same auditory stimulus engages broadly similar sensory and cognitive processing across trials, the resulting ERPs provide a common reference point for examining differences in brain activity across periods in which participants did and did not experience discomfort [8].

During ERP data analysis, data is segmented from the continuous raw EEG signal beginning from typically 100 to 200 ms prior to onset and ending one to two seconds after the stimulus. This time window of EEG data around the stimulus onset for a single stimulus (trial) is called an epoch. Changes in the signal at the trial and condition (sets

of trials within an experimental manipulation) level can reveal novel and consequential insights into the neural and cognitive processes engaged by the stimulus [8, 10, 11].

Throughout more than half a century of research, reliable ERP waveform features (components) have been mapped to cognitive states and resources. For example, the P3b component (also called the P3 or P300), is a positive-going potential that is typically maximal over centro-parietal electrodes beginning as early as 300 ms after an attended, task-relevant stimulus [12–16]. The amplitude of the P3b has consistently been shown to be influenced by task difficulty, cognitive load, and attention, and has been used in myriad studies to investigate attentional allocation [12, 13, 15]. One way to elicit a P3b is with an auditory oddball paradigm, wherein participants listen to a series of similar sounds (the standards). Randomly and infrequently, one of the sounds will be different (the oddballs), and participants must note these or make a response when they hear an oddball [12]. In previous work, we showed that changes in the auditory P3b were found to be associated with the intensity of cybersickness experienced [17]. However, there are likely other patterns present in the ERP data that reflect sickness and discomfort, including in response to the presentation of the standards, which elicit much less P3b activity. Uncovering these previously unknown discomfort-related patterns was the motivation for this work.

2.2 Cybersickness and Discomfort

When a person uses a VR head-mounted display, they sometimes develop feelings of discomfort, called cybersickness [2]. Cybersickness symptoms include nausea, dizziness, eyestrain, headache, and fatigue, among others [2, 18, 19]. If the sensations arise from using VR specifically, this discomfort can be referred to as VR sickness [1]. Theories on the causes for cybersickness include a conflict between the expectation of particular patterns of incoming sensory information and the patterns of sensory information experienced in that moment [20, 21], intravestibular imbalances [22], vestibulo-autonomic responses [23–25], head pose in the virtual environment [26], unexpected sensations of self-motion (vection) [27, 28], and others [1, 29].

There are a number of contributors to cybersickness. These include hardware, software, environmental, and individual factors [2, 18, 19, 29–31]. While not everyone that uses VR experiences discomfort or gets sick, a significant portion of users can develop adverse symptoms [29, 32, 33]. Although brief, these symptoms can sometimes last from 10 minutes to four hours after exposure in the HMD has ended [34].

There are different ways that cybersickness can be measured, but the most common method is to have research participants report how they are feeling after exposure with the Simulator Sickness Questionnaire (SSQ) [35]. The SSQ is comprised of 16 symptoms to which participants must give a rating from 0 (none of that symptom) to 3 (severe intensity of that symptom). Though several concerns about the SSQ have been raised [31], it is still widely used [2], which allows results across studies to be compared.

2.3 Cybersickness Recognition from EEG

EEG datasets are often small as a result of significant requirements in terms of time (EEG studies typically take around three hours per subject), resources (for the system

and per-run consumable costs), challenge (subject impacts like vomiting in the case of cybersickness studies), and other limitations (there is difficulty in recruiting a sample with high numbers of people who will get sick, as making the study extremely sickening poses risks in human subject research). As a result, EEG datasets, particularly for cybersickness studies, are quite limited, with one review finding that only 18% (6 total) of EEG cybersickness studies had more than 30 participants [3]. Moreover, there is substantial variability in EEG data, making it difficult for models to generalize.

Despite this challenge, other work has explored using machine learning for cybersickness detection from EEG [6, 36–38]. This usually involves feature extraction before training a classifier. Features extracted are often in the frequency or time-frequency domain relating to power in frequency bands [36, 38–42]. A combination of EEG features with other modalities like inertial measurement unit for head movement [40] or biosignals like heart rate, blink rate, and electrocardiogram [42] has also been explored. Some work extracts time domain features like variance, standard deviation, and number of peaks [37] or Principal Component Analysis (PCA) components [39]. Commonly chosen classifiers are k-Nearest Neighbors (kNN) [39, 41], Support Vector Machine (SVM) [39, 41], Naive Bayes (NB) [37], Linear Discriminant Analysis (LDA) [41], Convolutional Neural Network (CNN) [43], Long Short-Term Memory (LSTM) [40], AdaBoost [36, 37], and more. Another line of work directly interacts with EEG in the time domain without first extracting features, usually with neural networks [43–45]. Reported cybersickness detection accuracies can range around 76% [40] and 80% [36] with train and test sets separated across different subjects. However, the time window sizes vary from a few (three to six) seconds [36, 40] to longer windows like 30s [38]. In contrast, the data used in this work are periods of 800 ms after the sound of each auditory stimulus, a much shorter time window to use for identifying cybersickness discomfort. For evaluation, we note that leave-one-subject-out cross validation is a more rigorous method to determine how well a model can generalize to unseen subject data [40, 43] as opposed to mixing all subjects’ data together for dataset splits. Due to the usually small dataset size, it may be better to repeat experiments on all possible folds for analysis, rather than using a small k for k-folds. As our goal here is to identify interpretable EEG features for cybersickness discomfort classification, neural networks with an explainability method is our approach.

2.4 Complications in Classifying Cybersickness

One of the many complicating factors in elucidating neural activity associated with VR sickness arises from the way cybersickness is quantified with the SSQ. The measure was originally developed for pilot trainees using Navy helicopter flight simulators in the 1990’s [35], and, accordingly, the list of symptoms selected for inclusion in the SSQ reflects what most commonly arose when using those devices. As current HMDs bear little resemblance to the simulators from over 30 years ago, the symptoms experienced now may not be well-captured by the SSQ questionnaire [31]. Another complicating factor from the SSQ is precision in data labeling. VR users rate their symptoms after exposure has completed, meaning the data is given a cybersickness score as a batch although sickness intensity could have fluctuated during exposure. This motivates the need for real-time detection of cybersickness and related discomfort as it is occurring.

A novel contribution of this work is our exploration of brain activity associated with one cybersickness symptom, general discomfort, in ERP data. The 16 symptom severities included in the SSQ total weighted score span a variety of visual, physiological, and cognitive reactions that may each produce *independent* activation which passes through divergent areas of the brain [3]. This may explain why our prior work with deep learning on ERP cybersickness data labeled with SSQ scores was unable to achieve high classification accuracy [46]. If we use labels from the full SSQ or SSQ subscales, we are asking the models to parse multiple different streams of brain activity in a very small dataset, which could dilute our ability to identify sickness-relevant patterns. Here, by classifying data labeled with discomfort only, we can focus the models on patterns of brain activity related to an unpleasant aspect of cybersickness: discomfort.

2.5 EEG Analysis with ML Model Interpretability

Explainable artificial intelligence to understand ML model decision-making can benefit EEG analysis by identifying which features were most useful to push the model’s decision in the right direction. An ablation-style analysis can be done by modifying each part of the model and observing changes in accuracy metrics [47]. Since CNNs are often used in EEG analysis, Gradient-weighted Class Activation Maps (GradCAM), which extract activations in the feature map from a target convolutional layer, are a common choice [48, 49], often for EEGNet [50]. Saliency maps are another option, which outline features from the input that most contribute to the correct decision [43]. Integrated gradients that calculate the attribution of input features to model prediction is another choice [45]. With the introduction of transformers, especially vision transformer (ViT), attention maps reveal how different parts of an image (or EEG data shaped similarly to an image) influence the ViT’s decision [49]. While these features can be useful, as has been noted [6], care should be exercised in the interpretation of salient features identified by ML models. These features may be correlated with the focus of the classification, but other more highly correlated biomarkers may be overlooked by the models. More assessment is needed to verify the generalizability of the features, for example, with statistical tests and additional datasets. However, the ML models do point to a direction in which further study and investigation can be done.

3 Results

We share results from running our open-source framework that trains deep learning models and extracts model interpretability maps [code will be released on acceptance]. The framework was used to classify discomfort given a 0.8 second window of ERP data and extract which spatio-temporal features of the input data were most useful for the model’s correct classification decisions. Here, we report the findings from three different neural networks: a convolutional neural network EEGNet [50], a convolutional transformer EEG Conformer [51], and a modified ImageNet pre-trained vision transformer EEG ViT [52]. The classification is evaluated in two phases: 1) leave-one-subject-out (LOSO) cross validations to evaluate how well the models’ learned patterns generalize to unseen participants and 2) per-subject calibration to evaluate how well the models can learn patterns more specific to a participant, due to

high inter-subject variability in EEG. This evaluation is performed over two datasets. The results for the first ERPVR dataset are shared in Section 3.1. Results for the second VectionVR dataset are in Section 3.2. To interpret which data features contributed to the model for classification, we use integrated gradients [53] which assign an importance score to each feature in the input data. For EEGNet we also use GradCAM [54], which shows the activations in the model’s feature map that had more focus. More on model interpretability is in Section 7.7 and the Supplemental Material Section F.

3.1 ERPVR Dataset Results

3.1.1 ERPVR Model Accuracies

We show results from 120 experiment runs, varying over different configurations of the pipeline for the ERPVR dataset in Table 1. We define a sample as “discomfort” when the participant’s General Discomfort rating on the SSQ had a score > 0 and “no discomfort” as a General Discomfort score $= 0$. The configurations were permuted per Figure 5D for standards vs. oddballs ERP data type, under vs. oversampling, outlier detection on vs. no outlier detection, and electrode order alphabetical vs. spatial. For each model, we aggregate the results by pipeline configuration and by participant groups: subjects who felt discomfort during all trials, subjects who felt no discomfort during all trials, and subjects who felt both (mixed). On average, we see the participant-specific calibration helps boost balanced accuracy for all three groups in Table 1.

For the all trials discomfort participant group, for participant-specific calibration, ViT was able to achieve the highest balanced accuracy around 95.6% with oversampling, no outlier detection, and spatial electrode ordering. For base LOSO to evaluate generalization to unseen participants, EEGNet performed the best with balanced accuracy around 76.9% using Standard trials EEG data, oversampling, outlier detection, and alphabetical electrode ordering. ViT follows closely with balanced accuracy around 76.3% in base LOSO with Standard trials data, oversampling, outlier detection, and alphabetical electrode ordering. For the all trials no discomfort subjects, for subject-specific calibration, ViT also achieves the highest balanced accuracy around 89.6% with oversampling, no outlier detection, and spatial electrode ordering. Conformer follows closely around 89.3% with Standard trials data, undersampling, outlier detection, and spatial electrode ordering. For base LOSO, Conformer performed best at 82.6%. However, the mixed label subjects were more difficult for the models to classify, with the best average performance around 57.7% balanced accuracy with Conformer using Standard trials, undersampling, and no outlier detection. This may be due to intra-subject label differences, and although calibration still does help improve balanced accuracy, it is less for mixed label subjects compared to the single-label subjects, as seen in Table 1.

Even though mixed label subjects in general had lower balanced accuracies, there are a few where the model was able to achieve high balanced accuracy. They are especially interesting because these subjects indicate that some patterns learned by the models were able to classify and separate discomfort within that subject. One example is subject 15 (16.1% of trials are discomfort) with EEGNet, Standard trials, outlier detection,

M	D	S	O	E	DB	DC	NB	NC	MB	MC
EEG	Std	Und	ON	Alp	71.2 $\pm$ 7.0	83.6 $\pm$ 3.9	61.1 $\pm$ 2.3	67.9 $\pm$ 1.1	51.2 $\pm$ 0.7	51.1 $\pm$ 0.7
EEG	Std	Und	ON	Spa	73.0 $\pm$ 1.1	81.8 $\pm$ 1.8	58.9 $\pm$ 3.0	63.7 $\pm$ 1.9	52.3 $\pm$ 0.4	53.0 $\pm$ 1.7
EEG	Std	Over	ON	Alp	76.9 $\pm$ 2.0	84.8 $\pm$ 1.8	60.8 $\pm$ 2.6	65.9 $\pm$ 2.4	47.4 $\pm$ 2.7	47.1 $\pm$ 4.7
EEG	Std	Over	ON	Spa	71.2 $\pm$ 1.5	82.3 $\pm$ 4.3	58.0 $\pm$ 1.5	62.8 $\pm$ 3.1	50.8 $\pm$ 2.0	49.4 $\pm$ 4.6
EEG	Std	Und	OFF	Alp	47.2 $\pm$ 6.7	77.7 $\pm$ 6.5	76.4 $\pm$ 0.4	67.7 $\pm$ 4.4	51.7 $\pm$ 1.6	52.5 $\pm$ 1.7
EEG	Std	Und	OFF	Spa	48.8 $\pm$ 1.1	82.1 $\pm$ 3.7	75.2 $\pm$ 1.2	65.6 $\pm$ 4.6	51.7 $\pm$ 0.7	53.6 $\pm$ 0.7
EEG	Std	Over	OFF	Alp	49.4 $\pm$ 2.8	76.2 $\pm$ 2.0	76.9 $\pm$ 0.6	68.4 $\pm$ 5.2	51.7 $\pm$ 0.5	52.5 $\pm$ 1.4
EEG	Std	Over	OFF	Spa	50.8 $\pm$ 3.0	79.0 $\pm$ 3.0	76.0 $\pm$ 2.1	68.6 $\pm$ 0.8	51.4 $\pm$ 1.7	52.7 $\pm$ 2.1
EEG	Odd	Und	ON	Alp	75.8 $\pm$ 4.5	79.0 $\pm$ 5.8	57.7 $\pm$ 1.9	63.0 $\pm$ 0.3	52.0 $\pm$ 0.5	51.0 $\pm$ 0.5
EEG	Odd	Und	ON	Spa	73.9 $\pm$ 0.9	78.2 $\pm$ 1.4	56.7 $\pm$ 3.1	61.9 $\pm$ 4.0	51.9 $\pm$ 0.3	52.1 $\pm$ 1.8
EEG	Odd	Over	ON	Alp	73.7 $\pm$ 5.1	77.2 $\pm$ 3.2	58.6 $\pm$ 3.9	63.8 $\pm$ 3.5	50.9 $\pm$ 1.1	50.6 $\pm$ 1.5
EEG	Odd	Over	ON	Spa	71.6 $\pm$ 3.1	81.1 $\pm$ 1.7	52.9 $\pm$ 1.9	58.2 $\pm$ 2.3	50.7 $\pm$ 1.0	51.0 $\pm$ 2.3
EEG	Odd	Und	OFF	Alp	45.3 $\pm$ 3.0	62.8 $\pm$ 6.2	72.5 $\pm$ 3.0	70.2 $\pm$ 4.1	50.8 $\pm$ 1.1	51.9 $\pm$ 2.8
EEG	Odd	Und	OFF	Spa	50.0 $\pm$ 3.2	67.8 $\pm$ 5.7	73.4 $\pm$ 5.2	68.9 $\pm$ 4.4	50.6 $\pm$ 0.7	51.4 $\pm$ 0.2
EEG	Odd	Over	OFF	Alp	48.2 $\pm$ 1.9	68.2 $\pm$ 2.9	73.7 $\pm$ 4.5	68.9 $\pm$ 3.3	51.7 $\pm$ 0.8	51.8 $\pm$ 1.7
EEG	Odd	Over	OFF	Spa	53.2 $\pm$ 3.6	67.6 $\pm$ 5.0	77.3 $\pm$ 2.9	71.2 $\pm$ 5.3	51.7 $\pm$ 1.1	51.9 $\pm$ 0.7
Con	Std	Und	ON	Alp	51.9 $\pm$ 8.3	78.6 $\pm$ 4.6	76.7 $\pm$ 3.7	89.3 $\pm$ 2.7	52.5 $\pm$ 0.9	56.7 $\pm$ 2.9
Con	Std	Und	ON	Spa	57.3 $\pm$ 4.9	79.9 $\pm$ 3.1	75.7 $\pm$ 3.1	89.3 $\pm$ 2.6	53.8 $\pm$ 0.8	56.8 $\pm$ 0.7
Con	Std	Over	ON	Alp	50.1 $\pm$ 9.6	80.9 $\pm$ 3.6	76.7 $\pm$ 1.2	87.2 $\pm$ 2.2	51.5 $\pm$ 2.2	56.3 $\pm$ 2.6
Con	Std	Over	ON	Spa	49.7 $\pm$ 3.6	82.8 $\pm$ 2.8	73.8 $\pm$ 2.7	88.8 $\pm$ 1.4	51.7 $\pm$ 1.7	54.0 $\pm$ 2.2
Con	Std	Und	OFF	Alp	33.5 $\pm$ 6.7	84.6 $\pm$ 0.7	82.6 $\pm$ 0.4	85.9 $\pm$ 4.1	53.4 $\pm$ 1.2	57.7 $\pm$ 0.2
Con	Std	Und	OFF	Spa	33.2 $\pm$ 1.9	83.0 $\pm$ 3.7	81.8 $\pm$ 1.3	87.9 $\pm$ 4.8	53.2 $\pm$ 1.5	56.9 $\pm$ 1.1
Con	Std	Over	OFF	Alp	37.7 $\pm$ 2.5	84.3 $\pm$ 2.9	79.9 $\pm$ 0.9	86.5 $\pm$ 4.5	53.2 $\pm$ 1.8	56.7 $\pm$ 1.7
Con	Std	Over	OFF	Spa	34.5 $\pm$ 5.5	84.9 $\pm$ 1.4	82.3 $\pm$ 0.8	87.3 $\pm$ 3.3	54.0 $\pm$ 1.9	57.2 $\pm$ 0.6
Con	Odd	Und	ON	Alp	57.1 $\pm$ 2.7	69.7 $\pm$ 2.5	68.2 $\pm$ 0.6	74.1 $\pm$ 2.0	52.8 $\pm$ 1.3	52.3 $\pm$ 2.5
Con	Odd	Und	ON	Spa	59.4 $\pm$ 3.1	73.4 $\pm$ 7.4	67.6 $\pm$ 3.3	75.3 $\pm$ 2.5	53.6 $\pm$ 0.4	51.7 $\pm$ 1.4
Con	Odd	Over	ON	Alp	59.4 $\pm$ 1.5	69.4 $\pm$ 1.9	68.5 $\pm$ 2.1	75.6 $\pm$ 2.5	51.5 $\pm$ 1.7	52.8 $\pm$ 2.4
Con	Odd	Over	ON	Spa	56.0 $\pm$ 1.8	65.8 $\pm$ 5.1	71.2 $\pm$ 1.1	76.7 $\pm$ 3.5	53.2 $\pm$ 0.3	51.6 $\pm$ 1.8
Con	Odd	Und	OFF	Alp	35.0 $\pm$ 2.3	68.7 $\pm$ 10.5	78.4 $\pm$ 1.3	72.8 $\pm$ 2.8	53.3 $\pm$ 1.0	53.0 $\pm$ 0.9
Con	Odd	Und	OFF	Spa	30.9 $\pm$ 1.3	65.1 $\pm$ 9.2	75.0 $\pm$ 3.1	75.8 $\pm$ 2.8	51.3 $\pm$ 1.4	53.4 $\pm$ 1.9
Con	Odd	Over	OFF	Alp	32.7 $\pm$ 6.2	71.6 $\pm$ 3.9	76.0 $\pm$ 0.7	74.0 $\pm$ 4.3	52.2 $\pm$ 0.3	53.1 $\pm$ 2.1
Con	Odd	Over	OFF	Spa	32.2 $\pm$ 1.6	67.0 $\pm$ 5.3	76.7 $\pm$ 1.7	76.2 $\pm$ 0.2	50.8 $\pm$ 1.5	53.0 $\pm$ 1.3
ViT	Std	Und	ON	Alp	74.4 $\pm$ 2.2	92.9 $\pm$ 1.9	53.7 $\pm$ 5.6	86.6 $\pm$ 3.8	50.5 $\pm$ 0.7	54.0 $\pm$ 3.5
ViT	Std	Und	ON	Spa	72.6 $\pm$ 4.5	91.3 $\pm$ 0.7	50.4 $\pm$ 0.8	86.2 $\pm$ 1.6	51.4 $\pm$ 0.4	54.3 $\pm$ 0.5
ViT	Std	Over	ON	Alp	76.3 $\pm$ 0.3	92.3 $\pm$ 0.3	51.4 $\pm$ 0.7	87.1 $\pm$ 2.4	48.4 $\pm$ 2.6	53.1 $\pm$ 1.1
ViT	Std	Over	ON	Spa	71.2 $\pm$ 2.1	91.7 $\pm$ 0.6	54.3 $\pm$ 0.7	87.0 $\pm$ 4.0	51.2 $\pm$ 0.2	54.6 $\pm$ 0.9
ViT	Std	Und	OFF	Alp	51.0 $\pm$ 2.1	94.2 $\pm$ 1.6	65.6 $\pm$ 2.4	87.9 $\pm$ 1.9	51.6 $\pm$ 1.4	54.5 $\pm$ 1.9
ViT	Std	Und	OFF	Spa	49.8 $\pm$ 4.8	93.4 $\pm$ 2.0	66.8 $\pm$ 1.8	88.7 $\pm$ 3.2	52.8 $\pm$ 0.2	56.7 $\pm$ 1.3
ViT	Std	Over	OFF	Alp	48.7 $\pm$ 4.6	94.2 $\pm$ 1.1	65.8 $\pm$ 0.6	89.0 $\pm$ 1.8	51.9 $\pm$ 0.8	56.3 $\pm$ 1.0
ViT	Std	Over	OFF	Spa	45.7 $\pm$ 2.2	95.6 $\pm$ 0.3	67.7 $\pm$ 1.3	89.6 $\pm$ 1.8	51.6 $\pm$ 0.4	56.3 $\pm$ 0.9

Table 1: ERPVR dataset balanced accuracy from base LOSO and calibration phases by configuration $\times$ subject group averaged over random seeds. Each cell is run three times with different random seeds and averaged over subjects in the subject group to get one final average for that run. We report (mean $\pm$ (max – min)/2) over the three final averages. M is “Model” for model used: “EEG” for EEGNet, “Con” for EEG Conformer, and “ViT” for the pretrained EEG ViT. D is “Data type” for Standards or Oddballs. S is “Sampling” for Undersampling or Oversampling. O is “Outlier detection” turned on or off. E is “Electrode order” for alphabetical or spatial. DB is “Discomfort base”, DC is “Discomfort calibration”, NB is “No Discomfort base”, NC is “No Discomfort calibration”, MB is “Mixed base”, and MC is “Mixed calibration.”

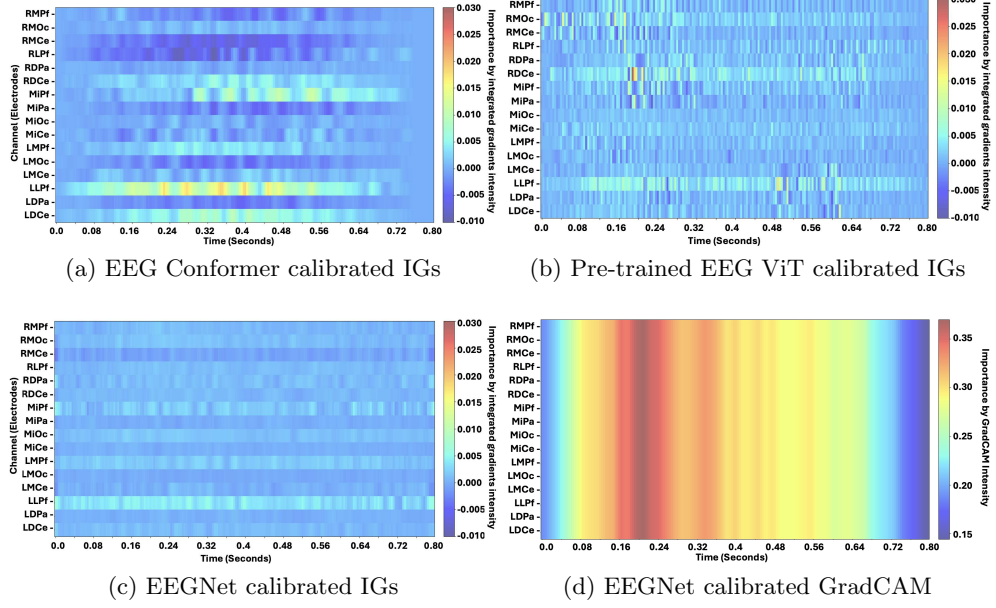


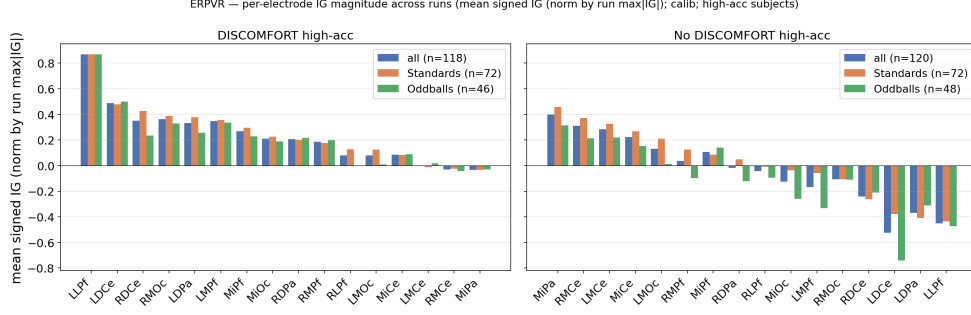
Fig. 2: ERPVR Integrated Gradients (IGs) averaged over sick data from subjects with balanced accuracy $> 75\%$. Warmer colors indicate features have more positive contribution for model classification. These plots are from one of the three model seeds.

and oversampling, which achieved 81.0% balanced accuracy after calibration. Another example is subject 10 (83.5% of trials are discomfort) with Conformer, Standard trials, outlier detection, and undersampling, which achieved 75.1% balanced accuracy after calibration.

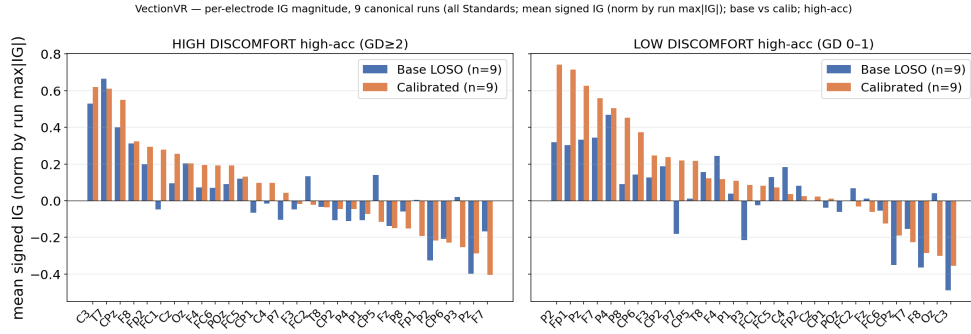
3.1.2 ERPVR Integrated Gradients

To determine which electrodes were most useful for discomfort classification, the integrated gradient (IG) for each electrode channel is summed over the time dimension. Then for each run, the summed IG is normalized by the max magnitude summed IG and averaged over the runs (see Figure 3a). The first plot shows results from runs where there are participants that experienced discomfort which achieved high balanced accuracy over 75%. The second plot is from runs where no discomfort subjects achieved high balanced accuracy. We also plot the frequency of an electrode reaching top-5 of the summed IG for a run (indicating that the electrode was one of the top five contribution electrodes for model classification) in Supplemental Figure B2.

For each run configuration, the integrated gradients (IGs) map for all models after calibration show the IGs averaged over discomfort data from high balanced accuracy ($> 75\%$) subjects. Figure 2 shows the results for Standard trials, outlier detection, oversampling, and alphabetical electrode ordering. Results for more runs, including trials when the models were trained in the spatial electrode ordering, are provided in the Supplemental Material. In terms of the most relevant features for the models, from the IG calculated across multiple runs, a prominent observation is that some



(a) ERPVR integrated gradients for each electrode summed over time, then normalized by the max summed IG magnitude per run, and averaged over all runs, using Standard trials only, or using Oddball trials only, after calibration. N is the number of experiment runs for each group. “All” combines all model runs, both with Standard trials and with Oddball trials. The ordering of electrodes shown is based on the importance calculations for the standards only. Numbers less than zero mean that electrode was negatively contributing for classification.



(b) VectionVR integrated gradients for each electrode summed over time then normalized by the max summed IG magnitude per run, and averaged over all runs for base LOSO vs. after calibration. These are Standard trial experiment runs only. High Discomfort is when “General Discomfort” (GD) is reported to be 2 or greater in the SSQ. Low Discomfort when “General Discomfort” is reported to be 0 or 1 in the SSQ. The ordering of electrodes shown is based on the importance calculations for calibrated.

Fig. 3: Integrated gradients for each electrode summed over time, then normalized by the max summed IG magnitude per run, and averaged over all runs. Top plot (a) is from ERPVR runs and bottom plot (b) is from VectionVR runs.

electrodes consistently had high integrated gradient positive attributions. LLPf had the highest integrated gradients for all models and runs, usually followed by LDCe and RDCe. RMOc and LDPa were also generally helpful for the models with a higher average sum of IG over the runs. In the time domain, in EEG Conformer’s IG map in Figure 2, for LLPf there were time steps highlighted between 240 to 440 ms, while

EEGNet’s GradCAM highlighted an earlier 160 to 240 ms window. Pre-trained EEG ViT highlighted around 160 to 200 ms for LLPf and RDCe.

Comparing the model interpretability maps from mixed label subjects that had balanced accuracy $> 75\%$, we may identify which features helped the model discriminate between discomfort and no discomfort trials within-subject. The integrated gradients highlight similar electrodes to ones highlighted by all discomfort subjects’ IG maps for contributing most to the model classifying discomfort: LDCe, LLPf, LMPf, RDCe, and MiPf. This may indicate that some of the models’ learned patterns can carry over to within-subject discomfort classification, as well as between-subject discomfort classification. The integrated gradients maps are in Supplemental Figure B1.

3.1.3 ERPVR Configurations

Based on the metrics increase post-calibration, it seems having some way to bridge the distribution shift between train set and test subjects, especially when working with a small dataset of complex and noisy data like EEG, is crucial for tailoring a classifier to an individual for cybersickness discomfort detection. Since the transformers performed better than EEGNet after calibration, this may indicate that the sequence modeling done by the transformers with attention in the encoder blocks is important for cybersickness discomfort detection in EEG data, which EEGNet as a CNN only, lacks.

For ERPVR, the initial ordering of channels was alphabetical, which did not impact performance for models that learned spatial filters across all electrodes. However, EEG ViT’s patch embedding is over four channels, so we re-ordered the channels to group electrodes based on their location spatially on the head and ran all models. We found the metrics were similar and the channels marked as important by IG remained consistent, though EEG ViT did see some changes in attribution. The electrode ordering did have a bigger effect for the VectionVR data runs however, discussed in section 3.2.

One configuration that performed well after calibration for discomfort classification for ERPVR was using Standard trials, oversampling, no outlier detection, and spatial ordering. With the ViT model, this achieved the highest average balanced accuracy for discomfort subjects with 95.6%. EEG Conformer also reached its best average performance with 84.9%. However, an interesting contrast is that this same configuration had worse performance in base leave-one-subject-out compared to other configurations. This configuration also performed well (compared to other configurations) for mixed label subjects for all three models. For no discomfort subjects, this same configuration also performed the best for ViT with 89.6% average balanced accuracy. However for EEG Conformer and EEGNet, they performed better with Standard trials, undersampling, and outlier detection with Conformer reaching around 89.3%.

3.1.4 ERPVR Statistical Analyses of Configuration Types

We used a Wilcoxon signed-rank test to test whether the different configurations for ERPVR runs had any significant differences, with detailed numbers reported in Supplemental Material Table B2. Since we ran multiple tests, we used Bonferroni Correction, setting $\alpha = \frac{0.05}{N} = 0.01$ for $N = 5$ statistical tests performed. We compared the pre-calibration base leave-one-subject-out phase and post-calibration and each of the four configurations: Standards vs. Oddballs data, outlier detection vs. no outlier

detection, over vs. undersampling, and alphabetical electrode order vs. spatial order. The Wilcoxon tests were performed by comparing runs where all configurations, except for the configuration being tested, are kept the same. Then, within the matched pair (e.g. Oversampling / Standards / Outlier detection / Alphabetical electrode order / EEGNet / seed 3 vs. Undersampling / Standards / Outlier detection / Alphabetical electrode order / EEGNet / seed 3 is one example of a pair if the configuration being tested is sampling type), we compute the difference $\Delta(\%)$ between each subject’s balanced accuracy in the two runs. The delta is computed by the first configuration - second configuration for each model run that satisfies the pairing. Then, this is averaged for each subject across all the paired runs, so we have one averaged delta per subject representing how much the configuration being tested affected the balanced accuracy for that subject. These deltas are then compared.

We did not find any significant differences between the configurations for outlier detection, sampling, data type, or channel order. However, there was a meaningful difference between calibration vs. base LOSO, with calibration positively increasing the balanced accuracy for participants. We also evaluated differences between each configuration choice with Wilcoxon signed-rank tests over balanced accuracies from the experiment runs for each of the three main subject groups in ERPVR: pure discomfort, mixed, and no discomfort subjects. These results are reported in Supplemental Material Section B.1.

3.2 VectionVR Dataset Results

We ran the main pipeline for VectionVR data as described in Section 7.5 on a second dataset, VectionVR, with results in Table 2. Due to the imbalance of the “General Discomfort” SSQ label in the VectionVR data, we separated the data into high discomfort (SSQ score for “General Discomfort” was 2 or 3) and low discomfort (SSQ score for “General Discomfort” was 1 or 0).

3.2.1 VectionVR Model Accuracies

We show results over nine experiment runs in Table 2, three random seeds for each of the three models, using the pipeline configuration as described in Section 7.5. We see that the base LOSO on average has much worse accuracy than for ERPVR runs, indicating that the VectionVR dataset was harder for the models to learn to generalize

Model	HD base	HD calib	LD base	LD calib	Overall base	Overall calib
EEGNet	49.9 $\pm$ 6.1	76.1 $\pm$ 4.6	52.9 $\pm$ 2.1	73.1 $\pm$ 3.7	51.6 $\pm$ 1.4	74.4 $\pm$ 1.4
Conformer	68.2 $\pm$ 5.3	88.6 $\pm$ 1.6	33.6 $\pm$ 1.8	82.6 $\pm$ 2.0	48.0 $\pm$ 2.0	85.1 $\pm$ 0.8
ViT	47.0 $\pm$ 4.1	89.9 $\pm$ 0.8	50.5 $\pm$ 3.4	92.6 $\pm$ 1.4	49.0 $\pm$ 3.3	91.5 $\pm$ 0.6

Table 2: VectionVR dataset accuracy averaged over random seeds. Each cell is run three times with different random seeds and averaged over subjects in the subject group to get one final average for that run. We report (mean $\pm$ (max - min)/2) over the three final averages and per subject group: High discomfort (“HD”) or Low discomfort (“LD”). Overall averages over all subjects. These accuracies are for the spatial channel order only.

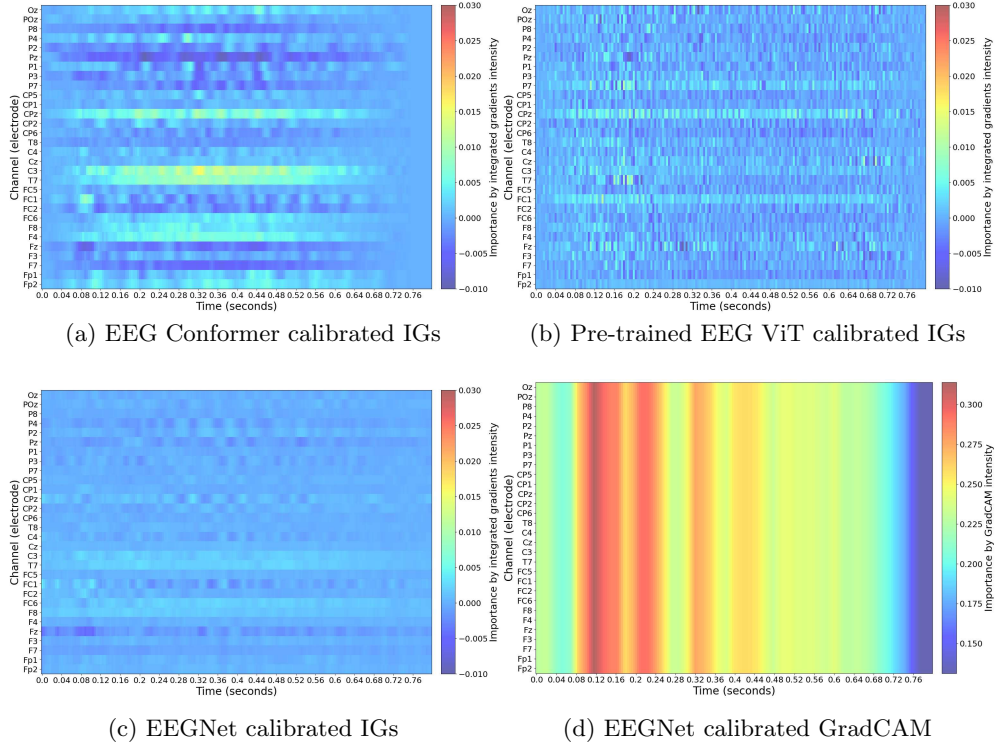


Fig. 4: VectionVR integrated gradients averaged over high discomfort data from subjects with balanced accuracy $> 75\%$ (from one run of the three seeds). Warmer colors indicate features have more positive contribution for model classification.

to unseen subjects. One possible reason could be due to the coarser resolution of the labels: ERPVR had four block-level SSQ labels while VectionVR had one label for all eight blocks. Calibration did help pull up the accuracy quite substantially for per-subject classification. Since we did not have mixed label subjects in VectionVR, we cannot evaluate within-subject high to low discomfort classification. And because there are only single label subjects, balanced accuracy collapses to regular accuracy.

From these runs, Imagenet pre-trained ViT performed the best after calibration with around 90% accuracy for high discomfort subjects, 92.6% for low discomfort subjects, and 91.5% overall. This indicates the calibration can help the model learn whether a subject is high or low discomfort, but not within-subject classification because of label resolution. For base leave-one-subject-out, EEG Conformer performed best for high discomfort while EEGNet performed best for low discomfort.

Some exploratory runs were used to test the effects of different configurations on VectionVR data. Outlier detection was tested at varying levels of strictness, and it was found that keeping outlier detection off (balanced accuracy averaged over all subjects in the run with base LOSO is around 52.6%) performed slightly better than with outlier detection on (49.9%). For sampling, oversampling (balanced accuracy around 50.9%) was found to be better than undersampling (around 40.6%). For electrode ordering, which specifically matters for the pre-trained ViT because the patch embedding groups

the channels together depending on the input order, spatial ordering was better at around 46.2% average balanced accuracy compared to alphabetical around 40.5%.

3.2.2 VectionVR Integrated Gradients

Similar to the analysis for ERPVR runs, we sum the integrated gradient of each electrode channel over the time dimension for the VectionVR data. Then for each run, the summed IG is normalized by the max magnitude summed IG and averaged over the runs (see Figure 3b). The left plot shows results from runs where there are high discomfort subjects that achieved accuracy over 75%, and the right plot for runs where there are low discomfort subjects that achieved accuracy over 75%. We also plot a histogram of the frequency of an electrode reaching top-5 summed IG for a run, indicating the electrode was a top-5 contributing electrode for model classification of discomfort in VectionVR, in Supplemental Figure B3.

Figure 4 shows the integrated gradients map for all models after calibration, where the IGs are averaged over high discomfort data from high balanced accuracy ($> 75\%$) subjects. This result is from Standard trials, oversampling, no outlier detection, and spatial electrode ordering. From these results, there are some electrodes that were consistently useful and positively contributed to high discomfort classification across the models. C3 and CPz were consistently useful for all three models. T7 contributed very positively for EEG Conformer and EEGNet, but not as much to pre-trained ViT. Similarly, F8 is useful for EEG Conformer and EEGNet, but relatively neutral for ViT. ViT highlights FC1 and P7 as useful electrodes for high discomfort classification, whereas they were pretty neutral for EEGNet and Conformer. This is also reflected in Figure 3b. We observe C3, T7, and CPz are the three most contributing spatial features over the runs.

In the time domain, EEGNet’s GradCAM in Figure 4 highlighted 80 to 240ms, which overlaps with the time window highlighted by EEGNet in ERPVR runs (from 160 to 200ms). EEG Conformer’s IG map highlighted a longer time window of 120 to 560ms. Pre-trained ViT seems to also highlight an early time window of around 80 to 280ms. The three models seem to overlap around the 120 to 240ms time window. This may indicate that there may be some patterns in the ERP data around this time window, especially since the data is time-locked, that help the models with discomfort classification. The windows may be slightly earlier for VectionVR compared to ERPVR as a result of differences in stimulus timing. For ERPVR, there was a small amount of latency with the auditory stimulus that was corrected with an auxiliary channel for time-locking during VectionVR data collection. Thus, the timing may be a bit more exact in the VectionVR data. More analysis plots in Supplemental Material Section B.

3.2.3 VectionVR Configurations

Similar to the ERPVR results, post-calibration helped increase metrics for VectionVR, reinforcing the importance of bridging the distribution shift between train set and test subjects for cybersickness discomfort detection for each subject. Unlike ERPVR, where the electrode order between alphabetical and spatial did not affect metrics too much, for VectionVR we found that spatial electrode ordering performed much better for base LOSO for the pre-trained ViT. This may be because of montage naming difference

between the two datasets. For ERPVR, alphabetical order sorts the electrodes into areas on the left and right sides of the head, but for VectionVR, which uses the 10-20 montage, alphabetical ordering sorts them roughly by region. Since ViT’s convolution is over a subset of electrodes at once, it is more likely to be affected by electrode channel ordering. Details about the model architectures in Supplemental Material Section E.

4 Discussion

In this article, we present a comprehensive evaluation of brain data collected while participants were experiencing cybersickness. We wanted to identify which temporal and spatial features were most helpful for classifying cybersickness-related discomfort in epoched ERP data collected from an auditory oddball task while participants viewed visual stimuli in VR. From datasets that were collected in different countries, with different VR environments, different VR headsets, different primary tasks, different EEG systems, and different intensities of cybersickness, we found similar patterns of important electrodes (left frontal, lateral central, and right parietal) and time periods (roughly around 80 to 260 ms post-stimulus). These patterns were highlighted by three different deep learning models across multiple random seeds and across multiple permutations of machine learning pipeline configurations. By training and evaluating the models on two completely separate datasets, we could assess whether patterns identified consistently in one dataset generalized to the other.

4.1 Important Spatial Features

Discomfort labels for the EEG data came from one item (general discomfort) of the SSQ questionnaire. The SSQ asks participants to rate at that moment whether their symptoms are none (0), slight (1), moderate (2), or severe (3). As a result of individual variability in cybersickness responses to the two different VR stimuli, the participant data split generally into three groups across the two (ERPVR and VectionVR) datasets: people who experienced no cybersickness-related discomfort, people who experienced mild-to-moderate levels of cybersickness discomfort, and people who experienced moderate-to-severe levels of cybersickness discomfort. For the no discomfort group in ERPVR, all ratings of general discomfort were 0. For the discomfort group in ERPVR, the ratings were primarily 1’s and 2’s, with an average score of 1.098. For the VectionVR data, the low discomfort group had three people with a score of 0, and the rest a score of 1, with an average of 0.857. For the high discomfort group, the scores were all 2’s and 3’s with an average of 2.133. Thus, the overlap between the two datasets happened in the discomfort ERPVR group and the low discomfort VectionVR group.

For participants who did not experience any cybersickness-related discomfort, centro-parietal regions (MiPa, LMCE, RMCE, MiCe, and LMOc) were the most useful features for the model. For participants that experienced intermediate levels of cybersickness discomfort (in both datasets), a distributed pattern was found across left frontal (LLPf, Fp1, F7) and right parietal (P2, P4, P8, RMOc) areas. Bilateral central (LDCe, RDCe) areas were helpful as well. For participants that experienced the highest levels of cybersickness discomfort, central and left centerline (C3, T7, CPz) areas were again tagged. Right frontal areas (F8, Fp2), to a lesser extent, were also useful.

Interesting patterns of specific important areas were highlighted despite the different layouts (montages) of electrodes for the two studies. For example, LDCe and C3 (and a little further out T7) along the left side of the head in central areas above the ears were tagged in both the ERPVR and VectionVR datasets as being useful for the group with a greater level of discomfort. LLPf and Fp1 with F7 (which are on prefrontal / frontal left areas of the head) as well as RMOc and P2 with Pz and P4 (which are all on right occipito-parietal areas of the head) were collectively determined as useful for the group that experienced discomfort in ERPVR and the low discomfort group in VectionVR. Overall, ERPVR generally elicited lower levels of discomfort than VectionVR, so this overlap may reflect participants experiencing broadly similar (intermediate) levels of discomfort across the two datasets. Another interesting pattern involved the MiPa and CPz electrode locations which are relatively close to each other. For the ERPVR dataset, MiPa was rated as the most important spatial feature for classification of the samples when people felt no discomfort. In the VectionVR dataset, CPz was the third most helpful feature for the samples with high discomfort. This similar location being important for both high and no discomfort could hypothetically be from high amplitudes in one case and low amplitudes in another, or the presence of signal in one case and absence of signal in the other. We cannot determine what exactly the model was picking up on during classification decisions. However, electrodes in centroparietal regions have been found as important features in other EEG cybersickness and frequency analysis studies [55], pain studies [56] and motion sickness studies [57].

The distributed pattern of informative scalp sites may also be expected given contemporary understanding of brain function. Complex perceptual, affective, and cognitive states arise through coordinated activity across distributed neural systems, with large-scale functional networks providing an organizing architecture that can be dynamically modulated by current task and state demands [58–60]. Cybersickness-related discomfort likewise likely reflects the combined activity of multiple sensory, autonomic, interoceptive, and affective areas. From this perspective, EEG is a particularly useful tool for investigating cybersickness because it samples electrophysiological activity simultaneously across the scalp, providing sensitivity to distributed networks and their rapidly changing patterns of neural activity [61]. Thus, the spatially distributed features identified by the models may reflect contributions from a broader network of brain regions involved in the experience of discomfort.

For ERPVR, across 120 model runs, with and without calibration, with and without outlier detection, with undersampling and oversampling, with Oddballs or Standards, and with two different channel orders, LLPf was consistently the channel with the highest integrated gradients, meaning that the left frontal scalp region helped the most in cybersickness discomfort classification. This result harmonizes with previous research and narrows in on an important component of cybersickness: the feeling of discomfort. Yang, Kasabov, and Cakmak [62] also found that activity in a left frontal electrode site above the inferior frontal gyrus was an important feature in predicting cybersickness. Other studies have found beta power in left frontal regions to be important as well [41, 63]. Thus, we postulate that tagging the left frontal region as a feature may be useful for (online) cybersickness discomfort detection.

It is important to note that signals from activity in different brain areas are distorted as they pass through the skull and meninges and propagate across the head. Additionally, the amplitudes that are picked up from an individual electrode can reflect the combination of activity from numerous neuronal ensembles, including those at some distance, making it difficult to map between (two-dimensional) scalp space and (three-dimensional) brain space (the EEG inverse problem [8]). This makes it very challenging to determine where in the brain the signals we record with EEG are coming from. It is also difficult to determine what it was about the patterns in the original signal that the models were picking up on after convolution. As such, care should be exercised in translating from ML model features to characteristics of underlying brain activity.

For this reason, we cannot interpret with certainty the neural or functional significance of the pattern in amplitudes that were here determined as an important feature by our models. At the same time, it is interesting to note what is known about the activity in left frontal and prefrontal areas from fMRI research (which provides substantially greater spatial specificity) regarding pain and sickness processing. Prior work has pointed to the prefrontal and orbitofrontal cortices as areas involved in regulation of autonomic outflow [64], and nausea induced fromvection stimulation during fMRI has been shown to result in activation in frontal areas [65]. Moreover, activity in the left dorsolateral prefrontal cortex (DLPFC) and left ventrolateral prefrontal cortex were significantly correlated with autonomic reactions (change in skin conductance) during recovery fromvection-induced nausea [64], further implicating vestibulo-autonomic responses in cybersickness [23–25].

Furthermore, the consequential role of the DLPFC in pain regulation and suppression, especially for headache, has been well-established [66–68]. Abnormalities in the DLPFC have been found in patients with chronic back pain [67] and migraines [69], and the relationship between migraines [68, 69] and cybersickness has been investigated previously [70]. Interestingly, transcranial magnetic stimulation (TMS) to the left DLPFC specifically has been demonstrated as an effective treatment for thermal pain [71] and migraine [69]. In view of this, it would be interesting to explore if TMS to the left DLPFC could reduce discomfort from cybersickness.

4.2 Important Time Features

Amplitudes recorded at the scalp are comprised of mixtures of activity generated by multiple neuronal populations over periods ranging from a few to hundreds of milliseconds. As a result, amplitudes at any specific time point or scalp location are highly correlated with activity at temporally and spatially adjacent points. Features identified as important by the models would therefore be expected to extend over ranges of time and space rather than appear at isolated points. In EEGNet’s GradCAM maps, importance in the ERPVR dataset increased beginning around 100 ms, peaked around 200 ms, and declined by approximately 300 ms. In VectionVR, the corresponding window began around 80 ms and extended to approximately 250 ms, with a peak around 110 ms. Although the precise timing shifted somewhat across seeds, the maps generally showed similar gradients. These patterns are broadly consistent with the timing of relatively early auditory sensory processing. Cortical auditory ERP components can be observed beginning within the first several tens of milliseconds after stimulus onset,

with prominent sensory responses appearing over approximately the first 200 ms [8]. These early responses are relatively tightly time-locked to the eliciting stimulus, whereas later activity increasingly reflects attentional, mnemonic, and other processes whose timing may vary across trials and individuals. One possibility, therefore, is that the relatively stereotyped timing of early auditory processing provides a temporally aligned signal in which differences associated with participants’ ongoing discomfort state are particularly accessible to the model.

The use of auditory event-related potentials during visual VR tasks was one contribution of this work, as most previous ML approaches to EEG-based cybersickness detection have used continuous EEG data [3, 6]. Repeated auditory events provide periods of neural processing that are precisely aligned in time and relatively constrained by a common eliciting stimulus. This consistency may help models identify variation associated with cybersickness-related discomfort that would be more difficult to isolate in unconstrained continuous EEG. Consistent with this possibility, EEGNet’s GradCAM maps in both datasets emphasized relatively early portions of the auditory-evoked response, despite substantial differences in the primary VR tasks performed in the two studies. In one case, participants were viewing a simulated museum and were asked to pay attention to the artwork and statues, about which they were quizzed at the end of the study. In the other experiment, participants were viewing rotating black and white lines and indicating when they felt sensations of self motion. In both cases, while participants did those primary tasks in VR, their secondary task was to listen to a series of beeps and respond by silently counting (ERPVR) or pressing a button (VectionVR) when they heard an infrequent random beep that sounded different from the others (the auditory oddball paradigm, see Section 2.1). The next question naturally becomes whether an auditory ERP task would be necessary to find similar features or use these features online. For practical purposes, it would be easier to not have VR users completing any task. If repeated time-locked auditory events are important for model sensitivity, a passive auditory paradigm, such as one designed to elicit the mismatch negativity (MMN; [72]), in which the person can ignore sounds playing in the background, could provide such probes without requiring an overt task. Alternatively, the spatial patterns identified here could guide analyses of continuous EEG to determine whether related discomfort-sensitive activity can be detected without any external probe.

4.3 Framework and Models

EEG data are multi-dimensional, complex, and variable across individuals, and EEG datasets are often small because of logistical constraints on data collection. Thus, we needed to develop a framework that would allow us to use different types of models, handle participants with imbalanced label data, calibrate the data by individual to overcome the issue of variability, and determine which features were most helpful for the models. Our framework, “Neural Network Interpretability with Leave-One-Subject-Out and subject-specific Calibration” (NNILC), is open-source and can be used by other researchers to classify ERP data with imbalanced labels and map out the importance of features to models. The method presented here utilizes CNNs and transformer models, calibration per data source, and interpretability maps to provide

a visual representation of what the models focused on during classification. NNILC is designed to be modular to allow evaluation at different phases of training and handle classification label imbalance, especially in the per-subject calibration. The modular design of the two phases, the base training with LOSO evaluation and later the per-subject calibration phase, is so the models can be evaluated on two different abilities: 1) generalization to new unseen subject data with the LOSO phase and 2) subject-specific calibration to learn patterns more specific to that subject. Neural networks were chosen to allow this two phase split of training from scratch (for compact models) and fine-tuning to a specific individual. NNILC is also model-agnostic; any neural network with classification capability can be loaded into this framework.

Using a sample of ERP brain activity collected while 29 participants were exposed to cybersickness-inducing stimuli in VR for ERPVR and 36 participants who were exposed to vection and cybersickness-inducing stimuli in VR for VectionVR, we compared the highest ranked features between models of increasing complexity by trainable parameters: a CNN (EEGNet), a convolutional transformer (EEG Conformer), and a pre-trained vision transformer (EEG ViT). We found that certain areas of the head (left frontal and right parietal) best helped predict cybersickness consistently. These features provide testable hypotheses about cybersickness-related patterns in EEG data and could be used to improve cybersickness classification in the future.

From the metrics, for ERPVR, calibration generally seemed to improve the models' classification performance per subject, though subjects with both labels were harder for the models. This may be the result of a larger distribution shift between these subjects and the others in the train set, where the majority had only one label. Reducing $L2_{SP}$ regularization may allow models to calibrate more to these subjects, though there is the trade-off of possibly losing helpful patterns learned from other subjects and overfitting. Alternatively, training the model on only one subject's data or supplementing with more data could help. However, logistical constraints limit both options. The amount of EEG data that can practically be collected from a single participant is typically limited for machine-learning applications. Also, since brain recordings are sensitive and potentially identifiable, they are not always made publicly available.

A review of 33 studies on brain activity during cybersickness found associated changes across broad areas of the head and in every frequency band [3]. This heterogeneous evidence made it difficult to predict which features might best help classify cybersickness. Here, we extracted features only from participants who experienced cybersickness-related discomfort and for whom the models achieved high classification accuracy. Though the ERPVR data collection study used a stimulus that caused cybersickness in a different sample previously, experimenters could not control how much discomfort participants experienced. Thus, from 29 participants in ERPVR, only 14 felt discomfort, and high accuracy classification was only achieved for a subset of those. Thus, our features from ERPVR are from a small number of people: five for Conformer, five for EEGNet, and six for pre-trained EEG ViT. Although this does decrease the amount of contributing data, we have greater confidence that the identified features actually reflect something about cybersickness discomfort and not other patterns present in the data. This may explain why a fewer number of features were especially important here versus in other work. Indeed, when we compared the

most useful features between high accuracy discomfort subjects and all subjects where more participants did not feel discomfort, the IGs both within multiple runs of the same neural network and between different neural networks were different. However for VectionVR, more subjects were able to achieve high accuracy for high discomfort classification by the models. Unlike ERPVR where only a handful of subjects felt discomfort, most of VectionVR’s subjects felt some level of discomfort. There were 15 high discomfort subjects and all 15 were able to achieve good accuracy after calibration: 9 for EEGNet, all 15 for Conformer, and all 15 for pre-trained EEG ViT.

5 Limitations

Though we have shown that our method can handle this highly variable data, the datasets are small for ML, even if that is expected for EEG. The number of people that experienced cybersickness discomfort in the ERPVR dataset was limited because individual differences in susceptibility meant that a random sample of participants would have significant variety in how much cybersickness or discomfort they experience during a study [31]. Though there were a small number of subjects who experienced cybersickness discomfort and had high accuracy, the question of cross-dataset generalizability is addressed with the second dataset, and we provide our code so other EEG cybersickness datasets can be compared. Another limitation is that across both montages there were few occipital electrodes, so we do not know if those areas of the head would have recorded brain data helpful in cybersickness discomfort classification. Also, the labels for the datasets are quite coarse since the SSQ was given to subjects after a block of trials (about every two minutes for four periods) in ERPVR, and only once at the end of eight blocks for VectionVR, though this is standard for cybersickness measurement [3]. The scoring options for discomfort in the SSQ are also only 0 - 3, which does not allow for fine-grained analysis based on the intensity of sickness. Because feature interpretation focused on participants for whom classification accuracy was $> 75\%$, the features identified here may not fully characterize the neural patterns present in participants who were more difficult for the models to classify.

6 Conclusion

In this paper, we identify features of EEG data that deep learning models consistently find most useful when classifying cybersickness-related discomfort in two datasets. By using NNILC, our model-agnostic framework to train neural networks, calibrating them by individual, and extracting interpretability maps of the models, we identified the features in the input EEG data that each neural network found useful. We demonstrate use of this framework with a CNN and transformers which classified whether or not a person was experiencing cybersickness discomfort given a short window of brain activity data. Across 120 ERPVR experiment runs varying trial type, sampling, outlier detection, and electrode order, a subset of scalp locations consistently contributed strongly to accurate discomfort classification. Related spatial and temporal patterns were also identified in the independent VectionVR dataset despite differences in participants, VR stimuli, EEG systems, and electrode montages. These recurring features provide testable

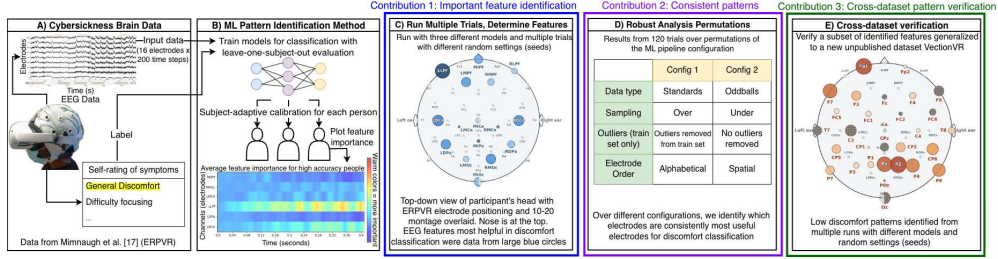


Fig. 5: Methods overview. A) ERPVR data from Mimnaugh et al. [17] was brain activity recorded with EEG while participants experienced cybersickness. B) Discomfort rating classification with leave-one-subject-out evaluation to measure model generalizability across subjects before per-subject calibration to evaluate subject-specific classification performance (further details in Figure 6). To ensure feature specificity, for people with discomfort classification accuracy $>75\%$, extracted feature interpretation maps from each model. C) Over multiple experiments, a subset of EEG electrodes consistently contributed the most information to the models for cybersickness-related discomfort classification. D) To verify the patterns, ablation experiments were run with all permutations of different configurations of the pipeline. E) Cross-dataset verification of identified patterns on a new, unpublished dataset (VectionVR).

candidates for future work, including studies examining whether similar patterns can support discomfort detection in continuous EEG without an auditory ERP paradigm.

7 Methods

7.1 ERPVR Dataset Description

This dataset is from Mimnaugh et al. [17]. The research study was conducted at the University of Illinois Urbana-Champaign and received approval from the IRB (#22519). Participants consented to data sharing and future use of their data for new research. The data were from 29 participants (17 female and 12 male, mean age 23.1).

ERPVR Study design. As described in the Supplemental Material Section K and original manuscript, ERPVR cybersickness data was collected using a dual-task auditory oddball paradigm. Participants passively viewed two tours through a virtual museum, each twice, in an Oculus Quest 2 headset (Figure 5A) while their brain activity was recorded. The two tours through the museum differed based on their amount of rotation and forward speed, and were shown in a previous study to evoke either mild or moderate amounts of VR sickness [73]. During the sickening museum tours, participants were instructed their primary task was to pay attention to the virtual museum because they would be quizzed on it later. They were also instructed to simultaneously listen to a series of beeps and keep a mental count of the infrequent (oddball) beeps (the P3b secondary task). After a non-sickening VR immersion condition (baseline) and immediately after each two minute tour through the VR museum, participants filled out an SSQ questionnaire rating their cybersickness symptoms during the museum

tour or at the current moment. All participant SSQ scores and other demographic and questionnaire responses are available from [17] online at: osf.io/v9fst.

ERPVR EEG data. Sixteen scalp electrodes, three electrooculogram, and two reference (left and right mastoids) were used for data collection. The full electrode montage was arranged in an approximately equidistant configuration to ensure uniform spatial sampling of the scalp potential field [74] (see [75] for a comparison to the 10-20 layout). A broadly distributed subset (16 of 32) electrodes was used to shorten data collection time (under three hours). For these analyses, ERP windows (epochs associated with each trial) of 800 ms after the participant heard each beep were extracted.

We separately ran experiments with the standard and oddball trials because there is significantly different cognitive processing in the ERP window related to making a response after the oddballs. As the dataset was already small and the ratio of standards to oddballs was about 70% to 30% per block (9,583 standards and 3,641 oddball total), we selected the standard trials for full analyses using all deep learning models. Selecting the standards provided the benefit that we could find patterns in the data not associated with the P3b, which we already knew to be impacted by cybersickness. Regardless of whether or not a participant responded to the oddball, cognitive processing of cybersickness-related discomfort (if experienced) was happening during both the oddball and standard trials. The oddball trials were run with only the more compact models that can handle the smaller dataset size.

ERPVR data preprocessing. Following guidance on ERP data preprocessing [8], raw EEG data were downsampled to 250 Hz, re-referenced to the average of both mastoids, cleaned of eye artifacts using independent component analysis, filtered, cut into periods of 800 ms after the auditory stimulus, baselined to the 200 ms before the stimulus, and then removed of any trials where voltages exceeded ± 100 microvolts from artifacts [17]. Per recommendations for free viewing EEG experiments [76] and optimal filter settings for the P3 [77], the EEG data was filtered with second-order IIR Butterworth filter with a 12 dB/octave roll-off and a 0.2 to 30 Hz half-amplitude cutoff.

ERPVR samples. There were approximately 330 Standard and 126 Oddball trials per person with 200 samples per trial in the time domain. Binary classification labels were based on each user’s self-reported intensity for one VR sickness symptom (general discomfort from the SSQ). If participants reported any level of discomfort (score > 0), then the sample was labeled as “discomfort,” else it was “no discomfort.” Binary classification was chosen instead of four-class because in the dataset only a subset of people felt discomfort (14 of 29), leaving few examples per class where the score > 0 . Therefore, the samples could not be further segmented into varying levels of discomfort.

7.2 VectionVR Dataset Description

The VectionVR dataset was collected at the John Paul II Catholic University of Lublin. Ethical approval for the project was obtained from the Research Ethics Committee of the Institute of Psychology (# KEBN_25/2024). Participants consented to future use of their data for new research. The data are from 36 of 40 participants (19 females and 17 males, mean age 22.7, min = 19, max = 39). Three of the 40 people were excluded (for atypical stereovision; an equipment issue; and vomiting), and one person did not consent to future use of their data.

VectionVR study design. Similar to ERPVR, this study was designed with a dual-task auditory oddball paradigm. However, there were several significant differences. This study was designed to determine how a factor that contributes to cybersickness, vection, impacted attention as compared to cybersickness. To elicit vection and cybersickness in participants, the subjects viewed alternating black and white lines which spun either clockwise or counterclockwise for each approximately two-minute block. The stimuli appeared similar to [78], except one outer ring of lines spun in one direction, and a smaller inner ring of lines spun in the opposite direction. Eight blocks were viewed in a VR HMD, the Meta Quest 3, and another eight blocks on a large screen (the screen trials were not included here). While viewing the lines, participants indicated if they were feeling vection by pressing on a joystick in their dominant hand, and responded to the oddball auditory stimulus with a button in their non-dominant hand. At the end of eight blocks, the participants completed an SSQ questionnaire. Thus, all eight blocks had only one discomfort rating for each subject.

VectionVR EEG data. A Brain Vision actiCHamp plus amplifier and actiCAP active electrodes with a standard 10-20 montage were used to collect EEG data. Thirty scalp electrodes and two electrodes removed from the EEG cap and affixed to the mastoids (TP9 and TP10) were used. The montage was modified from a 10-20 layout to provide a greater density of electrodes around centro-parietal areas. Data were recorded at 1000 Hz and impedances were kept under 10 k Ω . Windows (epochs) of EEG data for 800 ms after the participant heard the beep were extracted for analysis. Only standards were analyzed (as discussed in Section 2.1).

VectionVR data preprocessing. The VectionVR EEG data were preprocessed in EEGLAB [79] with the same preprocessing pipeline as the ERPVR data (Section 7.1, see [17] for a more detailed description of the processing). Data were recorded online with a Cz reference and re-referenced to the average of the mastoids offline. The data were downsampled to 250 Hz and filtered with a Butterworth bandpass filter (second order IIR, 0.2 to 30 Hz half-amplitude cutoff, 12 dB/octave roll-off). They were cleaned of artifacts using the same ICA method as ERPVR [76] and epoched from -200 ms pre-stimulus to 800 ms post-stimulus, with pre-stimulus baselining. A ± 100 μ V simple voltage threshold was used to remove epochs with artifacts in channels POz, CP1, Pz, P3, P1, P2, P4, CP2, CPz, and Cz (a similar area as ERPVR) across the one second window. For the models, screen trials and the 200 ms pre-stimulus period were removed.

VectionVR samples. There were approximately 336 Standard and 144 Oddball trials per person with 200 samples per trial in the time domain. EEG data were recorded on 30 channels (electrodes), giving the input data a size of 30 x 200. Binary classification labels were based on each user’s self-reported intensity for the same VR sickness symptom (general discomfort from the SSQ). However, unlike for ERPVR data, only three out of 36 subjects reported a score of 0. To handle this imbalance, we instead separate between low discomfort (score ≤ 1) and high discomfort (score > 1). This gave us 15 out of 36 high discomfort subjects and 21 low discomfort subjects. Since the SSQ was only collected once at the end of the block of HMD trials, there is only one SSQ label for each subject and not per-block like in ERPVR. So for VectionVR, there are only single-label subjects and no mixed label subjects. In total, there were 12,096 standard and 5,184 oddball trials.

7.3 Data Analysis and Augmentation

Machine learning (ML) models learn patterns within the train set to generalize to the validation set and unseen test set, so if the train set is skewed or biased in some way, the model could pick up that bias and learn something not inherent in the data. Since the ERPVR binary classification labels were very skewed (3:1 for no discomfort vs discomfort based on the SSQ symptom general discomfort), we analyzed the data through plots and unsupervised learning (in Supplemental Material Section D), and explored augmentation methods to alleviate this issue. This included balancing the labels through sampling (both over and undersampling), generating synthetic data, and identifying outliers in the data. More on data augmentation in Supplemental Material Section C.

Z-score. Input data size is a (C, T) matrix by channel (electrode) C and by time T . We first flattened the data to size $C * T$ before computing the z-score, defined as $\mathbf{z}_i = \frac{\mathbf{x}_i - \mu}{\sigma}$ where $\mathbf{x}_i \in \mathbb{R}^{C*T}$ is the flattened vector being normalized, $\mu \in \mathbb{R}^{C*T}$ is the mean, and $\sigma \in \mathbb{R}^{C*T}$ is the standard deviation, calculated differently to serve two distinct purposes with different computations.

1. Per-feature normalization: mean and standard deviation were computed per feature across the train set regardless of label. Each of the 3200 channel-time features for ERPVR and 6000 features for VectionVR has its own mean and standard deviation, thus channels are not forced to share a common distribution. This is necessary for neural network training stability. Without it, EEG’s varying amplitudes would cause gradient instability and bias the model toward larger absolute amplitude features regardless of discomfort content.
2. Outlier detection: mean and standard deviation were computed separately per label per feature from the train set. Trials where any feature exceeds $z=3$ are removed. This prevents extreme discomfort-related patterns from being flagged as outliers against the no discomfort distribution. Z-score ≥ 3 is used to delineate outliers in other work [80, 81]. We apply it at finer granularity per feature and validate by comparing strict, lenient, and no removal. Validation and test sets are never modified, so these steps cannot inflate evaluation metrics.

Outlier detection and removal. Adding outlier removal improved performance compared to no outlier removal for subjects with discomfort-related sickness labels for ERPVR, analysis in Supplemental Table B3. Since removing trials where any feature exceeds $z=3$ was quite strict, we also tested a milder version that identified outliers if over half the features exceeded three standard deviations. However, test performance dropped for ERPVR, especially on subjects with more discomfort data, compared to the strict version. For VectionVR, no outlier detection performed better than with outlier detection.

Sampling. To balance the class labels, we tried variations of under and oversampling. For ERPVR, for undersampling, we randomly picked majority no discomfort data to match the number of discomfort. For oversampling, we randomly replicated minority discomfort data until the classes were balanced. For VectionVR, it was also high discomfort as the minority class so oversampling would sample from there, and low discomfort is the majority class so undersampling would drop from that.

These approaches were not ideal, however, since with undersampling we lose data, whereas with oversampling repeating the same data does not provide new information to the models. Generating synthetic data for the minority class was investigated but found to perform worse than random sampling for neural network models.

7.4 Main Data Processing Pipeline for ERPVR dataset

The best performing data processing pipeline for ERPVR data classification of discomfort was: 1) individual stimuli outlier detection and removal, 2) normalization, and 3) random oversampling. The strict outlier detection was run first on the train set only, then z-score normalization on the remaining data. Validation, test, and calibration sets were normalized with the same average μ and standard deviation σ obtained from the train set. Finally, the train set was augmented using random oversampling, implemented with Pytorch’s [82] weighted random sampler by giving higher weights to the minority class data. To check for consistency however, we vary the random seed so the sampler picks different samples. We separate the standards and oddball ERPs because the cognitive processing is different between the two.

7.5 Main Data Processing Pipeline for VectionVR dataset

The best performing data processing pipeline for VectionVR data classification of high vs. low discomfort was: 1) no outlier detection, 2) normalization, and 3) random oversampling. Spatial electrode channel ordering does better than alphabetical for VectionVR’s 10-20 montage, especially for the ImageNet pre-trained Vision Transformer model (more in Section 7.6). The normalization and random oversampling use the same method as in ERPVR, z-score normalization by average μ and standard deviation σ from train set and Pytorch’s weighted random sampler.

7.6 Deep Learning for Pattern Recognition

Since unsupervised learning by clustering the EEG data without labels was unable to identify meaningful groupings, and traditional machine learning models struggled to capture the complexities of the EEG data in the time domain (more in Supplemental Material Section D) , deep learning models were considered. Convolution seemed promising to extract spatially local features in the signal, so EEGNet [50] was chosen as it was designed for EEG data, including ERPs, and compact enough to work with small datasets. Since EEG has a time dimension component, models that take sequential information into account were also considered, specifically transformers. EEG Conformer was chosen as it is a compact convolutional transformer designed to encapsulate local and global features in EEG data [51]. An ImageNet pre-trained EEG Vision Transformer (ViT) was chosen to leverage the reasoning patterns it had learned from images to EEG data with fine-tuning [52]. Each input sample is a matrix of electrical potential amplitudes in microvolts with shape EEG electrodes. For ERPVR it is $C = 16$ by time $T = 800\text{ms}$ (200 time steps in the data) and for VectionVR it is $C = 30$ by time $T = 800\text{ms}$ for 200 time steps as well. The three models are described by order of increasing complexity.

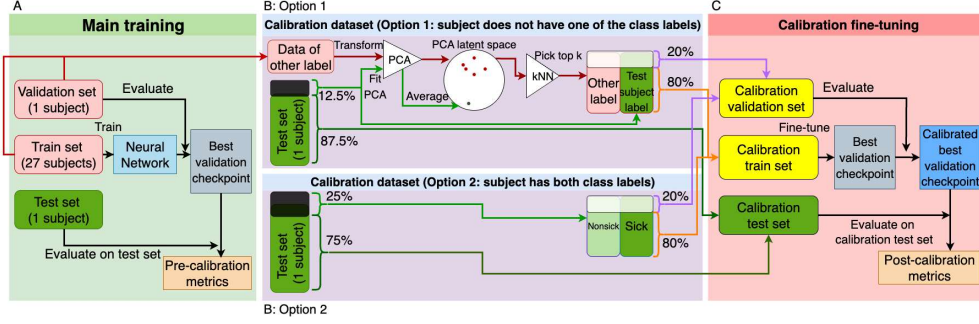


Fig. 6: Neural Network Interpretability with Leave-One-Subject-Out and subject-specific Calibration (NNILC) Framework: 1) main training (A) for leave-one-subject-out cross validation and 2) calibration (B and C) to fine-tune model to the test subject. There are two possible calibration datasets depending on if the test subject data has both labels or not. For main training, the train set goes through outlier detection, z-score normalization, and random oversampling. Best validation checkpoint is fine-tuned in calibration. If test subject has both labels, 25% of the test subject’s data is sampled. If test subject only has one label, 12.5% of test subject’s data is sampled and top k data of the other label from other subjects are taken that are closest to sampled test subject data centroid in whitened PCA space.

7.7 Model Interpretability

For EEGNet, the final convolution layer was set for GradCAM extraction. The heatmaps show where along the time dimension the activation was strongest for the model to classify discomfort. This is because EEGNet’s final convolution layer is a pointwise convolution that mixes all feature maps together. For EEG Conformer and pre-trained EEG ViT, integrated gradients (IG) provides more fine-grained information about the gradient strength by channel and time. While GradCAM interpretation very much depends on which layer the maps were extracted from, IG passes through the whole model to compute the cumulative gradients. Descriptions in Supplemental Material F.

7.8 Calibration

Due to the high inter-subject variability of the ERP data, it was difficult for the model to generalize to the data of the subject left out. To bridge the domain shift between the train set distribution and the subject left out’s distribution, with the goal of improving subject-specific accuracy, we used some data of the left out subject to fine-tune the trained model checkpoint (Figure 6). Many works on EEG classification tasks rely on calibration to specific subjects to improve classification performance, by collecting labeled data of the test subject to build a classifier [83, 84] or fine-tune a model [85].

Calibration dataset. We sampled either 12.5 or 25% from the dataset of the subject left out. If the subject left out had both labels, then 25% of the data were sampled, where exactly 12.5% were from each class. However, many subjects only had one label, so for those subjects, 12.5% of their data was sampled. To handle the imbalance, the same amount of the other class label were taken from the remaining 28

subjects' data for ERPVR or 35 for VectionVR. Instead of random sampling, we took the top k -nearest neighbors that were closest to the subject left out's distribution since the goal of the calibration set was to bridge the distribution shift for each individual subject.

The subject left out's sampled data were used to fit a whitened PCA space of 16 components for both ERPVR and VectionVR. Each sampled data was projected to this space then averaged into a centroid. Next, each of the other label data from the remaining subjects was projected into PCA space and kNN was run to find and retrieve the data that were closest to the centroid by Euclidean distance. K was the same as 12.5% of the subject left out's data. Together, this formed the calibration dataset.

The remaining test subject's data were used as a test set. The validation set was formed by taking 20% of each class from the calibration dataset. Since each subject had about 330 data samples for Standards, about 40 for each class label were sampled for calibration for ERPVR. For VectionVR Standards, each subject has 336 samples, about 42 for each class label were sampled for calibration. If the test subject only had one label, then about 40 of the other label was taken from the other subjects for a total of 80 data samples, of which approximately 16 were used as validation for ERPVR. For VectionVR, since all subjects only had one label, then about 42 of the other label was taken from the other subjects for a total of 84 data samples, of which approximately 18 were used as validation.

Calibration fine-tuning. Once the calibration dataset was created, the model with the best validation score from the main training phase was selected to be fine-tuned by the calibration set. The validation metric for this phase was a custom weighted balanced accuracy to identify early stopping from evaluation on the validation set, and was designed to have more weight on the correct classification of discomfort data.

For EEGNet, all layers were unfrozen to fine-tune all weights. For the EEG Conformer and pre-trained EEG ViT, sections of the model were frozen. For EEG Conformer, the first three of four encoder blocks were frozen. For the pre-trained EEG ViT, all 12 encoder blocks remained frozen. L2-starting point ($L2_{SP}$) [86] regularization, defined in Equation 1 where w_0 is a vector of model parameters from the best validation checkpoint after the main training phase and w is the current model's unfrozen parameters in the calibration fine-tuning phase and β is the regularization strength, was added to prevent the pre-trained weights from drifting too much during fine-tuning. This was crucial to prevent the model checkpoints from overfitting to the small calibration set and "forgetting" the distribution it had learned during the main training phase. In experiment runs, $\beta = 2 \times 10^{-3}$. The L2-starting point regularization is defined as:

$$L2_{SP}(w) = \frac{\beta}{2} ||w - w_0||_2^2 \quad (1)$$

where S is the test subject's dataset, $b_i = \begin{cases} 2.0 & \text{if } i \in S \\ 1.0 & \text{otherwise} \end{cases}$ is the weight for data sample, $i \in N$ for calibration train set size N , y_i is the ground truth label, x_i is the classification output logits, and p_i is the softmax probability for the discomfort label 1 Equation 2, defined as:

$$p_i = \frac{e^{x_{i,1}}}{e^{x_{i,0}} + e^{x_{i,1}}}, \quad (2)$$

and the cross entropy loss (CE) is defined as:

$$CE(x, y) = - \sum_{i=1}^N \frac{b_i}{\|\mathbf{b}\|} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)). \quad (3)$$

We used cross entropy loss (CE) with normalized weight by sample and $L2_{SP}$ regularization to form the calibration phase loss here:

$$L(x, y, w) = CE(x, y) + L2_{SP}(w). \quad (4)$$

Custom weighted balanced accuracy. Since we placed more emphasis on the model’s ability to correctly classify discomfort data at the cost of potentially creating more false positives, we created a custom metric to identify an early train stopping point by validation. This custom weighted balanced accuracy (WBA) allows tuning the weight on the recall of the discomfort class (positive label) and no discomfort class (negative label) while keeping the weights normalized. Let TP be true positive, TN true negative, FP false positive, and FN false negative. In experiment runs, $\alpha = 0.7$ for the weighted balanced accuracy, defined as:

$$\text{WBA} = \alpha \left(\frac{TP}{TP + FN} \right) + (1 - \alpha) \left(\frac{TN}{TN + FP} \right). \quad (5)$$

7.9 Experiment Setup and Metrics

For ERPVR, each model runs each configuration of the pipeline for three trials with different random seeds. For VectionVR, each model runs the main pipeline for three trials with different random seeds. Exact training, calibration, and testing details in Supplemental Material Section G.

To evaluate our approach in the main training pre-calibration phase, we use “leave-one-subject-out” cross-validation where one specific subject was left out of the train set to serve as the test set and another subject is randomly chosen to serve as the validation set. The remaining subjects (27 for ERPVR, 34 for VectionVR data) were in the train set. This ensured that the test set was completely unseen by the model at this phase and the evaluation results reflected how well the model could generalize to the unseen subject. For the calibration phase, the validation and test sets were as described in Figure 6 where minimal amounts of the test subject’s data is seen by the model, with the evaluation reflecting how well the model closed the distribution shift to the test subject. Since the dataset was very imbalanced, the metric we considered is balanced accuracy. Another metric for imbalanced data, F1-score, is reported in Supplemental Material Section A.

Balanced accuracy. Balanced accuracy (BA) is the average of the recall for each class. Let TP be true positive, TN true negative, FP false positive, and FN false negative. Balanced accuracy for binary classification [87] is:

$$\text{BA} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right). \quad (6)$$

7.10 LLM Usage

AI was used to create the montage images (Figure 1) from our electrode location coordinates and calculated IGs, plot integrated gradients charts, format tables in latex, find related literature, and proofread the final draft.

Supplementary information. This article does have supplementary files.

Acknowledgements. The authors wish to thank Ruohan Zhang, Dariusz Zapala, Hanna Zgorzelska, Weronika Mroczka, Markku Suomalainen, Israel Becerra, Rafael Murrieta-Cid, Eliezer Lozano, Larry, Lucy, Yoshi, Abdulsatar Aboud, Michael Mimnaugh, James Motes, and Marco Morales.

Declarations

8 Funding

This work was supported by a European Research Council Advanced Grant ILLUSIVE: Foundations of Perception Engineering, 101020977.

9 Data availability

Framework code and additional analyses will be available on Github. ERPVR participant data, with the exception of raw EEG (for participant privacy protection), is available on OSF (<https://osf.io/v9fst>). Aggregate EEG data measurements for ERPVR are on Harvard Dataverse (<https://doi.org/10.7910/DVN/MENRFT>). VectionVR data, with the exception of raw EEG, will be made available on OSF on publication of the paper analyzing its ERPs.

10 Code availability

Framework code will be available on Github on acceptance.

11 Author contribution

K.M., J.Y., K.F., S.L., T.O., W.Y., P.A., P.S., and N.A. contributed to the conception. K.M., J.Y., E.C., K.F., P.A., P.S., and M.K. contributed to the design. J.Y. and K.M. wrote the paper. J.Y., K.M., K.F., M.K., E.C., and P.A. contributed to the analysis. K.F., K.M., J.Y., E.C., and M.K. contributed to the interpretation of the results.

12 Competing interests

The authors declare no competing interests.

References

- [1] LaValle, S.M.: Virtual Reality. Cambridge university press, (2023)

- [2] Chang, E., Kim, H.T., Yoo, B.: Virtual reality sickness: A review of causes and measurements. *International Journal of Human–Computer Interaction* **36**(17), 1658–1682 (2020) <https://doi.org/10.1080/10447318.2020.1778351>
- [3] Chang, E., Billinghurst, M., Yoo, B.: Brain activity during cybersickness: A scoping review. *Virtual Reality* (2023) <https://doi.org/10.1007/s10055-023-00795-y>
- [4] Bastuji, H., Frot, M., Perchet, C., Magnin, M., Garcia-Larrea, L.: Pain networks from the inside: Spatiotemporal analysis of brain responses leading from nociception to conscious perception. *Human Brain Mapping* **37**, 4301–4315 (2016) <https://doi.org/10.1002/hbm.23310>
- [5] Garcia-Larrea, L., Bastuji, H.: Pain and consciousness. Elsevier Inc. (2018). <https://doi.org/10.1016/j.pnpbp.2017.10.007>
- [6] Yang, A.H.X., Kasabov, N., Cakmak, Y.O.: Machine learning methods for the study of cybersickness: a systematic review. *Brain Informatics* **9**(1) (2022) <https://doi.org/10.1186/s40708-022-00172-6>
- [7] Yildirim, C.: A review of deep learning approaches to eeg-based classification of cybersickness in virtual reality. *IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR 2020*, 351–357 (2020) <https://doi.org/10.1109/AIVR50618.2020.00072>
- [8] Luck, S.J.: *An Introduction to the Event-related Potential Technique*, 2nd edn. MIT Press, (2014)
- [9] Keshavarz, B., Riecke, B.E., Hettinger, L.J., Campos, J.L.: Vection and visually induced motion sickness: How are they related? *Frontiers in Psychology* **6**, 1–11 (2015) <https://doi.org/10.3389/fpsyg.2015.00472>
- [10] Donchin, E.: Surprise!...surprise? *Psychophysiology* **18**, 493–513 (1981) <https://doi.org/0048-5772/81/050493-21>
- [11] Woodman, G.F.: A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception & Psychophysics* **72**(8), 2031–2046 (2010) <https://doi.org/10.3758/APP.72.8.2031>
- [12] Polich, J.: Updating p300: An integrative theory of p3a and p3b. *Clinical Neurophysiology* **118**(10), 2128–2148 (2007) <https://doi.org/10.1016/j.clinph.2007.04.019>
- [13] Ghani, U., Signal, N., Niazi, I.K., Taylor, D.: Erp based measures of cognitive workload: A review. *Neuroscience and Biobehavioral Reviews* **118**(July), 18–26 (2020) <https://doi.org/10.1016/j.neubiorev.2020.07.020>
- [14] Verleger, R.: Effects of relevance and response frequency on p3b amplitudes:

Review of findings and comparison of hypotheses about the process reflected by p3b. *Psychophysiology* **57**, 1–22 (2020) <https://doi.org/10.1111/psyp.13542>

- [15] Vanbilsen, N., Kotz, S.A., Rosso, M., Leman, M., Triccas, L.T., Feys, P., Moumdjian, L.: Auditory attention measured by eeg in neurological populations: systematic review of literature and meta-analysis. *Scientific Reports* **13**(1), 1–14 (2023) <https://doi.org/10.1038/s41598-023-47597-5>
- [16] Ladouce, S., Donaldson, D.I., Dudchenko, P.A., Ietswaart, M.: Mobile eeg identifies the re-allocation of attention during real-world activity. *Scientific Reports* **9**, 15851 (2019) <https://doi.org/10.1038/s41598-019-51996-y>
- [17] Mimnaugh, K.J., Center, E.G., Suomalainen, M., Becerra, I., Lozano, E., Murrieta-Cid, R., Ojala, T., LaValle, S.M., Federmeier, K.D.: Virtual Reality Sickness Reduces Attention During Immersive Experiences. *IEEE Transactions on Visualization and Computer Graphics* **29**(11), 4394–4404 (2023) <https://doi.org/10.1109/TVCG.2023.3320222>
- [18] Caserman, P., Garcia-Agundez, A., Gámez Zerban, A., Göbel, S.: Cybersickness in current-generation virtual reality head-mounted displays: systematic review and outlook. *Virtual Reality* **25**(4), 1153–1170 (2021) <https://doi.org/10.1007/s10055-021-00513-6>
- [19] Saredakis, D., Szpak, A., Birkhead, B., Keage, H.A.D., Rizzo, A., Loetscher, T.: Factors associated with virtual reality sickness in head-mounted displays: A systematic review and meta-analysis. *Frontiers in Human Neuroscience* **14**(March) (2020) <https://doi.org/10.3389/fnhum.2020.00096>
- [20] Reason, J., Brand, J.J.: Motion Sickness. Academic Press, . (1975). <https://psycnet.apa.org/record/1976-12574-000>
- [21] Reason, J.T.: Motion sickness adaptation: A neural mismatch model. *Journal of the Royal Society of Medicine* **71**(Nov), 819–829 (1978)
- [22] Previc, F.H.: Intravestibular balance and motion sickness. *Aerospace Medicine and Human Performance* **89**(2), 130–140 (2018) <https://doi.org/10.3357/AMHP.4946.2018>
- [23] Bogle, J.M., Benarroch, E., Sandroni, P.: Vestibular-autonomic interactions: Beyond orthostatic dizziness. *Curr. Opin. Neurol.* **35**(1), 126–134 (2022) <https://doi.org/10.1097/WCO.0000000000001013>
- [24] Cohen, B., Dai, M., Yakushin, S.B., Cho, C.: The neural basis of motion sickness. *Journal of Neurophysiology* **121**(3), 973–982 (2019) <https://doi.org/10.1152/jn.00674.2018>
- [25] Muth, E.R.: Motion and space sickness: Intestinal and autonomic correlates.

- Autonomic Neuroscience: Basic and Clinical **129**(1–2), 58–66 (2006) <https://doi.org/10.1016/j.autneu.2006.07.020>
- [26] Palmisano, S., Allison, R.S., Kim, J.: Cybersickness in head-mounted displays is caused by differences in the user’s virtual & physical head pose. *Frontiers in Virtual Reality* **1**, 587698 (2020) <https://doi.org/10.3389/frvir.2020.587698>
 - [27] Teixeira, J., Miellet, S., Palmisano, S.: Unexpected vection exacerbates cybersickness during hmd-based virtual reality. *Frontiers in Virtual Reality* **3**, 1–14 (2022) <https://doi.org/10.3389/frvir.2022.860919>
 - [28] Teixeira, J., Miellet, S., Palmisano, S.: Effects of vection type and postural instability on cybersickness. *Virtual Reality* **28**(2), 82 (2024) <https://doi.org/10.1007/s10055-024-00969-2>
 - [29] Stanney, K., Lawson, B.D., Rokers, B., Dennison, M., Fidopiastis, C., Stoffregen, T., Weech, S., Fulvio, J.M.: Identifying causes of and solutions for cybersickness in immersive technology: Reformulation of a research and development agenda. *International Journal of Human-Computer Interaction* **36**(19), 1783–1803 (2020) <https://doi.org/10.1080/10447318.2020.1828535>
 - [30] Biswas, N., Mukherjee, A., Bhattacharya, S.: “are you feeling sick?” – a systematic literature review of cybersickness in virtual reality. *ACM Computing Surveys* **56**(11), 1–38 (2024) <https://doi.org/10.1145/3670008>
 - [31] Tian, N., Lopes, P., Boulic, R.: A review of cybersickness in head-mounted displays: Raising attention to individual susceptibility. *Virtual Reality* **26**(4), 1409–1441 (2022) <https://doi.org/10.1007/s10055-022-00638-2>
 - [32] Garrido, L.E., Frías-Hiciano, M., Moreno-Jiménez, M., Cruz, G.N., García-Batista, Z.E., Guerra-Peña, K., Medrano, L.A.: Focusing on cybersickness: pervasiveness, latent trajectories, susceptibility, and effects on the virtual reality experience. *Virtual Reality* **26**(4), 1347–1371 (2022) <https://doi.org/10.1007/s10055-022-00636-4>
 - [33] Szpak, A., Michalski, S.C., Saredakis, D., Chen, C.S., Loetscher, T.: Beyond feeling sick: The visual and cognitive aftereffects of virtual reality. *IEEE Access* **7**, 130883–130892 (2019) <https://doi.org/10.1109/ACCESS.2019.2940073>
 - [34] Dużmańska, N., Strojny, P., Strojny, A.: Can simulator sickness be avoided? a review on temporal aspects of simulator sickness. *Frontiers in Psychology* **9**(NOV) (2018) <https://doi.org/10.3389/fpsyg.2018.02132>
 - [35] Kennedy, R.S., Lane, N.E., Berbaum, K.S., Lilienthal, M.G.: Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology* **3**(3), 203–220 (1993) https://doi.org/10.1207/s15327108ijap0303_3

- [36] Li, X., Zhu, C., Xu, C., Zhu, J., Li, Y., Wu, S.: Vr motion sickness recognition by using eeg rhythm energy ratio based on wavelet packet transform. *Computer methods and programs in biomedicine* **188**, 105266 (2020)
- [37] Mawalid, M.A., Khoirunnisa, A.Z., Purnomo, M.H., Wibawa, A.D.: Classification of eeg signal for detecting cybersickness through time domain feature extraction using naïve bayes. In: *International Conference on Computer Engineering, Network and Intelligent Multimedia*, pp. 29–34 (2018). <https://doi.org/10.1109/CENIM.2018.8711320>
- [38] Rosanne, O., Benesch, D., Kratzig, G., Paré, S., Bolt, N., Falk, T.H.: To pre-process or not to pre-process? on the role of eeg enhancement for cybersickness characterization and the importance of amplitude modulation features. *Frontiers in Virtual Reality* **6**, 1–26 (2025) <https://doi.org/10.3389/frvir.2025.1468971>
- [39] Yu, Y.-H., Lai, P.-C., Ko, L.-W., Chuang, C.-H., Kuo, B.-C., Lin, C.-T.: An eeg-based classification system of passenger’s motion sickness level by using feature extraction/selection technologies. In: *International Joint Conference on Neural Networks*, pp. 1–6 (2010). <https://doi.org/10.1109/IJCNN.2010.5596739>
- [40] Demirel, B.U., Dogan, A.H., Rossie, J., Möbus, M., Holz, C.: Beyond subjectivity: Continuous cybersickness detection using eeg-based multitaper spectrum estimation. *IEEE Transactions on Visualization and Computer Graphics* **31**(5), 2525–2534 (2025) <https://doi.org/10.1109/TVCG.2025.3549132>
- [41] Khoirunnisaa, A.Z., Pane, E.S., Wibawa, A.D., Purnomo, M.H.: Channel selection of eeg-based cybersickness recognition during playing video game using correlation feature selection (cfs). *International Conference on Biomedical Engineering: Smart Technology for Better Society*, 48–53 (2018) <https://doi.org/10.1109/IBIOMED.2018.8534877>
- [42] Nam, Y., Kim, Y., Kim, H., Ko, H., Park, K.: Automatic detection of nausea using bio-signals during immersing in a virtual reality environment. *International Conference of the IEEE Engineering in Medicine and Biology* **2**, 2013–2015 (2001)
- [43] Borra, D., Magosso, E., *et al.*: Deep learning-based eeg analysis: investigating p3 erp components. *Journal of Integrative Neuroscience* **20**(4), 791–811 (2021)
- [44] Lee, Y., Alamaniotis, M.: Unsupervised eeg cybersickness prediction with deep embedded self organizing map. In: *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 538–542 (2020). <https://doi.org/10.1109/BIBE50027.2020.00093>
- [45] Liu, M., Yang, B., Xu, M., Zan, P., Chen, L., Xia, X.: Exploring quantitative assessment of cybersickness in virtual reality using eeg signals and a cnn-eca-lstm network. *Displays* **81**, 102602 (2024)

- [46] Aboud, A.: Identifying cybersickness features from eeg data using deep learning. Master’s thesis, University of Oulu (2023). <https://oulurepo.oulu.fi/handle/10024/46862>
- [47] Mortaga, M., Brenner, A., Kutafina, E.: Towards interpretable machine learning in eeg analysis. *Studies in Health Technology and Informatics* **283**, 32–38 (2021) <https://doi.org/10.3233/SHTI210538>
- [48] Yan, Y., Zhou, H., Huang, L., Cheng, X., Kuang, S.: A novel two-stage refine filtering method for eeg-based motor imagery classification. *Frontiers in neuroscience* **15**, 657540 (2021)
- [49] Shimizu, D., Watanabe, K., Dengel, A.: Feature estimation of global language processing in eeg using attention maps. In: *Proceedings of the Asian Conference on Computer Vision*, pp. 1706–1722 (2024)
- [50] Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering* **15**(5), 056013 (2018) <https://doi.org/10.1088/1741-2552/aace8c>
- [51] Song, Y., Zheng, Q., Liu, B., Gao, X.: Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **31**, 710–719 (2023) <https://doi.org/10.1109/TNSRE.2022.3230250>
- [52] Yang, R., Modesitt, E.: Vit2eeg: leveraging hybrid pretrained vision transformers for eeg data. *arXiv preprint arXiv:2308.00454* (2023)
- [53] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) *International Conference on Machine Learning. Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, (2017). <https://proceedings.mlr.press/v70/sundararajan17a.html>
- [54] Gildenblat, J., contributors: PyTorch library for CAM methods. GitHub (2021)
- [55] Yang, A.H.X., Kasabov, N., Cakmak, Y.O.: Machine learning methods for the study of cybersickness: a systematic review. *Brain Informatics* **9**(1), 1–21 (2022)
- [56] Rockholt, M.M., Kenefati, G., Doan, L.V., Chen, Z.S., Wang, J.: In search of a composite biomarker for chronic pain by way of eeg and machine learning: where do we currently stand? *Frontiers in Neuroscience* **17** (2023) <https://doi.org/10.3389/fnins.2023.1186418>
- [57] Li, Z., Zhao, L., Chang, J., Li, W., Yang, M., Li, C., Wang, R., Ji, L.: Eeg-based evaluation of motion sickness and reducing sensory conflict in a simulated autonomous driving environment. In: *2022 44th Annual International Conference*

- of the IEEE Engineering in Medicine & Biology Society (EMBC), vol. 2022-July, pp. 4026–4030. IEEE, (2022). <https://doi.org/10.1109/EMBC48229.2022.9871407> . <https://ieeexplore.ieee.org/document/9871407/>
- [58] Park, H.-J., Friston, K.: Structural and functional brain networks: From connections to cognition. *Science* **342** (2013) <https://doi.org/10.1126/science.1238411>
 - [59] Sporns, O.: Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience* **15**, 247–262 (2013) <https://doi.org/10.31887/DCNS.2013.15.3/osporns>
 - [60] Bressler, S.L., Menon, V.: Large-scale brain networks in cognition: emerging methods and principles. *Trends in cognitive sciences* **14**(6), 277–290 (2010)
 - [61] O’Neill, G.C., Tewarie, P., Vidaurre, D., Liuzzi, L., Woolrich, M.W., Brookes, M.J.: Dynamics of large-scale electrophysiological networks: A technical review. *Neuroimage* **180**, 559–576 (2018)
 - [62] Yang, A.H.X., Kasabov, N.K., Cakmak, Y.O.: Prediction and detection of virtual reality induced cybersickness: a spiking neural network approach using spatiotemporal eeg brain data and heart rate variability. *Brain Informatics* **10**(1) (2023) <https://doi.org/10.1186/s40708-023-00192-w>
 - [63] Oh, H., Son, W.: Cybersickness and its severity arising from virtual reality content: A comprehensive study. *Sensors* **22**(4), 1–26 (2022) <https://doi.org/10.3390/s22041314>
 - [64] Sclocco, R., Kim, J., Garcia, R.G., Sheehan, J.D., Beissner, F., Bianchi, A.M., Cerutti, S., Kuo, B., Barbieri, R., Napadow, V.: Brain circuitry supporting multi-organ autonomic outflow in response to nausea. *Cerebral Cortex* **26**(2), 485–497 (2016) <https://doi.org/10.1093/cercor/bhu172>
 - [65] Napadow, V., Sheehan, J.D., Kim, J., Lacount, L.T., Park, K., Kaptchuk, T.J., Rosen, B.R., Kuo, B.: The brain circuitry underlying the temporal evolution of nausea in humans. *Cerebral Cortex* **23**(4), 806–813 (2013) <https://doi.org/10.1093/cercor/bhs073>
 - [66] Apkarian, A.V., Bushnell, M.C., Treede, R.D., Zubieta, J.K.: Human brain mechanisms of pain perception and regulation in health and disease. *European Journal of Pain* **9**(4), 463 (2005) <https://doi.org/10.1016/j.ejpain.2004.11.001>
 - [67] Ong, W.Y., Stohler, C., Herr, D.: Role of the prefrontal cortex in pain processing. *Molecular Neurobiology* **56**(2), 1137–1166 (2019) <https://doi.org/10.1007/s12035-018-1130-9>
 - [68] Seminowicz, D.A., Moayedi, M.: The dorsolateral prefrontal cortex in acute and

- chronic pain. *Journal of Pain* **18**(9), 1027–1035 (2017) <https://doi.org/10.1016/j.jpain.2017.03.008>
- [69] Zhou, J., Wang, Y., Luo, X., Fitzgerald, P.B., Cash, R.F.H., Fitzgibbon, B.M., Che, X.: Revisiting the effects of rtms over the dorsolateral prefrontal cortex on pain: An updated systematic review and meta-analysis. *Brain Stimulation* **17**(4), 928–937 (2024) <https://doi.org/10.1016/j.brs.2024.07.011>
 - [70] Carvalho, G.F., Mehnert, J., Basedau, H., Luedtke, K., May, A.: Brain processing of visual self-motion stimuli in patients with migraine: An fmri study. *Neurology* **97**(10), 996–1006 (2021) <https://doi.org/10.1212/WNL.00000000000012443>
 - [71] Brighina, F., De Tommaso, M., Giglia, F., Scalia, S., Cosentino, G., Puma, A., Panetta, M., Giglia, G., Fierro, B.: Modulation of pain perception by transcranial magnetic stimulation of left prefrontal cortex. *Journal of Headache and Pain* **12**(2), 185–191 (2011) <https://doi.org/10.1007/s10194-011-0322-8>
 - [72] Sussman, E.S., Chen, S., Sussman-Fort, J., Dinces, E.: The five myths of mmn: Redefining how to use mmn in basic and clinical research. *Brain Topography* **27**, 553–564 (2014) <https://doi.org/10.1007/s10548-013-0326-6>
 - [73] Becerra, I., Suomalainen, M., Lozano, E., Mimnaugh, K.J., Murrieta-Cid, R., LaValle, S.M.: Human perception-optimized planning for comfortable vr-based telepresence. *IEEE Robotics and Automation Letters* **5**(4), 6489–6496 (2020) <https://doi.org/10.1109/LRA.2020.3015191>
 - [74] Oostenveld, R., Praamstra, P.: The five percent electrode system for high-resolution eeg and erp measurements. *Clinical Neurophysiology* **112**(4), 713–719 (2001) [https://doi.org/10.1016/S1388-2457\(00\)00527-7](https://doi.org/10.1016/S1388-2457(00)00527-7)
 - [75] Ganis, G., Kutas, M., Sereno, M.I.: The search for “common sense”: An electrophysiological study of the comprehension of words and pictures in reading. *Journal of Cognitive Neuroscience* **8**(2), 89–106 (1996) <https://doi.org/10.1162/jocn.1996.8.2.89>
 - [76] Dimigen, O.: Optimizing the ica-based removal of ocular eeg artifacts from free viewing experiments. *NeuroImage* **207**(November 2019), 116117 (2020) <https://doi.org/10.1016/j.neuroimage.2019.116117>
 - [77] Zhang, G., Garrett, D.R., Luck, S.J.: Optimal filters for erp research ii: Recommended settings for seven common erp components. *Psychophysiology* **61**(6), 1–24 (2024) <https://doi.org/10.1111/psyp.14530>
 - [78] Stróżak, P., Augustynowicz, P., Ratomska, M., Francuz, P., Fudali-Czyż, A.: Vection attenuates n400 event-related potentials in a change-detection task. *Perception* **48**, 702–730 (2019) <https://doi.org/10.1177/0301006619861882>

- [79] Delorme, A., Makeig, S.: Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods* **134**(1), 9–21 (2004)
- [80] Nolan, H., Whelan, R., Reilly, R.B.: Faster: Fully automated statistical thresholding for eeg artifact rejection. *Journal of Neuroscience Methods* **192**(1), 152–162 (2010)
- [81] Wu, W., Keller, C.J., Rogasch, N.C., Longwell, P., Shpigel, E., Rolle, C.E., Etkin, A.: Artist: A fully automated artifact rejection algorithm for single-pulse tms-eeg data. *Human Brain Mapping* **39**(4), 1607–1625 (2018) <https://doi.org/10.1002/hbm.23938> <https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.23938>
- [82] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library (2019). <https://arxiv.org/abs/1912.01703>
- [83] Shi, K., Huang, R., Li, Z., Lyu, J., Zhao, Y., Song, G., Cheng, H., Zhang, J.: Plug-and-play multi-domain fusion adaptation for cross-subject eeg-based motor imagery classification. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 9475–9481 (2025). IEEE
- [84] Huang, X., Xu, Y., Hua, J., Yi, W., Yin, H., Hu, R., Wang, S.: A review on signal processing approaches to reduce calibration time in eeg-based brain–computer interface. *Frontiers in Neuroscience* **Volume 15 - 2021** (2021) <https://doi.org/10.3389/fnins.2021.733546>
- [85] Ding, Y., Udompanyawit, C., Zhang, Y., He, B.: Eeg-based brain-computer interface enables real-time robotic hand control at individual finger level. *Nature Communications* **16**(1), 5401 (2025)
- [86] Xuhong, L., Grandvalet, Y., Davoine, F.: Explicit inductive bias for transfer learning with convolutional networks. In: International Conference on Machine Learning, pp. 2825–2834 (2018). PMLR
- [87] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [erpvrvectiondataanalysisissupplementarymaterialv4.pdf](#)