

Supplementary Material

Appendix A Additional Results

F1-score. F1-score is the harmonic mean between precision and recall [1] where TP is true positive, FP is false positive, and FN is false negative. F1-score is defined as:

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \quad (A1)$$

where the F1-scores for the ERPVR runs are reported in Supplementary Table A1. For single-label subjects, F1-score becomes degenerate and defaults this to $F1 = \frac{2 * TP}{2 * TP + FN}$. F1-scores not reported for VectionVR due to degeneracy from single-label subjects only.

Appendix B Additional Analyses

In addition to the normalized IG averaged across ERPVR runs, we also plot the frequency counts of when each electrode was top-5 most useful for an experiment run in Figure B2.

For VectionVR, this is in Figure B3. We also include the integrated gradient maps of ERPVR mixed label subjects that had balanced accuracy $> 75\%$ in Figure B1. Additional integrated gradient plots for VectionVR run with seed 3 in Figure B4b.

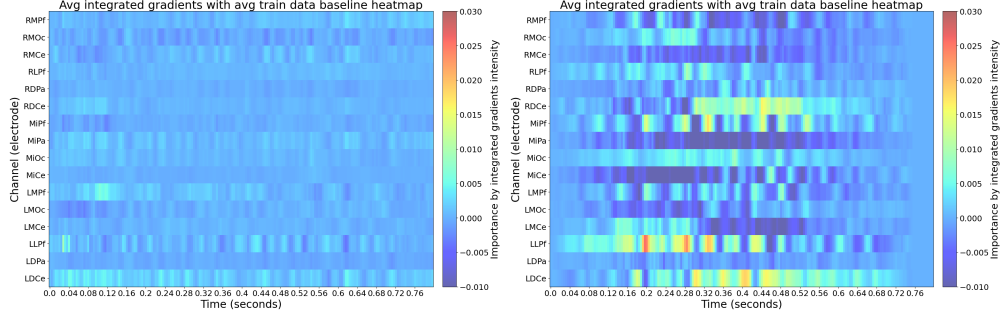
B.1 ERPVR Statistical Tests and Analyses

We report Wilcoxon signed-rank tests in Table B2 over all subjects in ERPVR. We also report Wilcoxon signed-rank tests in Table B3 by subject group subsets for ERPVR runs: subjects with all discomfort data, subjects with mixed labels, and subjects with all no discomfort data. They are treated as three separate families of five comparisons. Since we run $N = 5$ tests, the Bonferroni Correction is $\alpha = \frac{0.05}{5} = 0.01$ for statistical significance.

For the subject group split of the Wilcoxon tests in Table B3, for per-subject calibration compared to leave-one-subject-out base, we see that for all three subject groups, calibration led to higher balanced accuracies. For outlier detection, we see that for all discomfort subjects, having outlier detection on is helpful for increasing balanced accuracy. However, for all no discomfort subjects, it's in the opposite direction where having outlier detection decreased balanced accuracy. For sampling method, it seems over vs. undersampling had no statistical significance for all three subject groups. For data type of Oddballs vs. Standards, we do see for no discomfort subjects, there is

M	D	S	O	E	DB	DC	NB	NC	MB	MC
EEG	Std	Und	ON	Alp	79.2 $\pm$ 4.9	90.7 $\pm$ 2.4	71.3 $\pm$ 2.2	79.1 $\pm$ 1.3	45.8 $\pm$ 0.7	47.5 $\pm$ 3.6
EEG	Std	Und	ON	Spa	81.0 $\pm$ 1.0	89.6 $\pm$ 1.5	69.6 $\pm$ 2.8	75.7 $\pm$ 2.3	46.4 $\pm$ 0.9	47.5 $\pm$ 5.1
EEG	Std	Over	ON	Alp	84.4 $\pm$ 0.7	91.5 $\pm$ 1.2	70.9 $\pm$ 1.8	77.8 $\pm$ 1.7	44.2 $\pm$ 2.6	49.0 $\pm$ 3.7
EEG	Std	Over	ON	Spa	79.1 $\pm$ 1.0	89.6 $\pm$ 3.3	67.9 $\pm$ 1.5	75.6 $\pm$ 2.7	45.8 $\pm$ 1.3	47.1 $\pm$ 3.3
EEG	Std	Und	OFF	Alp	60.1 $\pm$ 6.0	87.0 $\pm$ 4.3	83.8 $\pm$ 0.6	79.2 $\pm$ 3.0	44.3 $\pm$ 0.9	50.2 $\pm$ 1.9
EEG	Std	Und	OFF	Spa	61.3 $\pm$ 1.7	89.9 $\pm$ 2.5	82.6 $\pm$ 1.0	77.4 $\pm$ 3.4	42.7 $\pm$ 1.2	48.4 $\pm$ 2.2
EEG	Std	Over	OFF	Alp	62.7 $\pm$ 2.5	86.2 $\pm$ 1.4	84.1 $\pm$ 0.9	79.8 $\pm$ 3.7	43.0 $\pm$ 0.5	49.2 $\pm$ 2.9
EEG	Std	Over	OFF	Spa	63.4 $\pm$ 4.0	88.0 $\pm$ 1.9	83.7 $\pm$ 2.1	80.0 $\pm$ 0.4	42.2 $\pm$ 2.4	50.6 $\pm$ 3.0
EEG	Odd	Und	ON	Alp	83.6 $\pm$ 3.4	87.0 $\pm$ 4.4	68.7 $\pm$ 2.3	74.8 $\pm$ 0.3	46.8 $\pm$ 2.7	49.4 $\pm$ 6.0
EEG	Odd	Und	ON	Spa	83.3 $\pm$ 0.7	87.2 $\pm$ 0.7	67.9 $\pm$ 3.1	73.6 $\pm$ 4.0	44.3 $\pm$ 1.5	47.5 $\pm$ 2.5
EEG	Odd	Over	ON	Alp	81.7 $\pm$ 3.6	85.2 $\pm$ 2.9	70.1 $\pm$ 3.2	75.4 $\pm$ 2.3	43.3 $\pm$ 1.1	45.0 $\pm$ 4.1
EEG	Odd	Over	ON	Spa	80.5 $\pm$ 2.0	89.2 $\pm$ 1.2	64.5 $\pm$ 1.7	70.8 $\pm$ 2.7	43.4 $\pm$ 1.9	44.4 $\pm$ 2.9
EEG	Odd	Und	OFF	Alp	59.2 $\pm$ 2.0	75.5 $\pm$ 5.7	80.6 $\pm$ 1.5	80.6 $\pm$ 3.5	42.7 $\pm$ 1.0	47.7 $\pm$ 7.4
EEG	Odd	Und	OFF	Spa	63.2 $\pm$ 1.7	80.2 $\pm$ 4.1	82.1 $\pm$ 3.7	79.9 $\pm$ 3.2	41.9 $\pm$ 0.3	46.9 $\pm$ 3.2
EEG	Odd	Over	OFF	Alp	60.1 $\pm$ 3.2	80.7 $\pm$ 2.0	81.5 $\pm$ 3.3	79.4 $\pm$ 2.6	41.7 $\pm$ 0.8	45.8 $\pm$ 3.4
EEG	Odd	Over	OFF	Spa	65.8 $\pm$ 4.2	79.7 $\pm$ 4.1	84.6 $\pm$ 1.8	80.8 $\pm$ 3.5	42.9 $\pm$ 2.5	47.4 $\pm$ 1.8
Con	Std	Und	ON	Alp	63.8 $\pm$ 7.7	87.5 $\pm$ 3.1	84.8 $\pm$ 3.7	94.2 $\pm$ 1.5	47.9 $\pm$ 1.3	59.8 $\pm$ 2.5
Con	Std	Und	ON	Spa	68.1 $\pm$ 4.2	88.5 $\pm$ 2.1	83.9 $\pm$ 3.2	94.2 $\pm$ 1.5	49.0 $\pm$ 2.8	58.9 $\pm$ 0.3
Con	Std	Over	ON	Alp	62.2 $\pm$ 7.8	89.2 $\pm$ 2.2	85.0 $\pm$ 1.4	92.9 $\pm$ 1.2	50.4 $\pm$ 2.5	60.2 $\pm$ 2.5
Con	Std	Over	ON	Spa	62.1 $\pm$ 2.4	90.3 $\pm$ 1.7	82.2 $\pm$ 2.3	93.9 $\pm$ 0.9	46.9 $\pm$ 2.3	57.1 $\pm$ 3.3
Con	Std	Und	OFF	Alp	47.0 $\pm$ 7.1	91.5 $\pm$ 0.5	89.0 $\pm$ 0.7	91.9 $\pm$ 2.6	47.0 $\pm$ 2.3	62.0 $\pm$ 0.8
Con	Std	Und	OFF	Spa	46.5 $\pm$ 1.5	90.5 $\pm$ 2.3	88.5 $\pm$ 1.5	93.2 $\pm$ 3.0	48.0 $\pm$ 2.5	59.8 $\pm$ 1.3
Con	Std	Over	OFF	Alp	51.1 $\pm$ 2.6	91.2 $\pm$ 1.9	87.2 $\pm$ 1.1	92.5 $\pm$ 2.8	47.0 $\pm$ 2.4	59.7 $\pm$ 2.1
Con	Std	Over	OFF	Spa	46.9 $\pm$ 4.8	91.6 $\pm$ 0.9	88.8 $\pm$ 0.7	93.0 $\pm$ 2.0	46.7 $\pm$ 1.5	59.4 $\pm$ 2.5
Con	Odd	Und	ON	Alp	69.5 $\pm$ 2.7	81.3 $\pm$ 2.1	78.4 $\pm$ 0.8	84.1 $\pm$ 1.7	47.5 $\pm$ 1.4	55.2 $\pm$ 3.3
Con	Odd	Und	ON	Spa	71.4 $\pm$ 2.9	83.8 $\pm$ 5.5	78.5 $\pm$ 2.3	85.2 $\pm$ 1.9	47.4 $\pm$ 0.8	55.5 $\pm$ 1.2
Con	Odd	Over	ON	Alp	71.0 $\pm$ 1.2	80.9 $\pm$ 1.1	78.9 $\pm$ 1.3	85.5 $\pm$ 1.8	46.3 $\pm$ 1.8	55.2 $\pm$ 1.4
Con	Odd	Over	ON	Spa	67.5 $\pm$ 1.5	77.8 $\pm$ 4.1	81.3 $\pm$ 1.0	86.3 $\pm$ 2.4	47.3 $\pm$ 2.0	54.8 $\pm$ 2.1
Con	Odd	Und	OFF	Alp	47.7 $\pm$ 2.9	79.5 $\pm$ 9.2	86.2 $\pm$ 1.5	83.2 $\pm$ 2.1	44.7 $\pm$ 1.8	51.7 $\pm$ 1.4
Con	Odd	Und	OFF	Spa	43.8 $\pm$ 0.8	76.7 $\pm$ 8.7	83.9 $\pm$ 1.9	85.6 $\pm$ 2.0	44.0 $\pm$ 1.1	55.1 $\pm$ 1.8
Con	Odd	Over	OFF	Alp	45.6 $\pm$ 6.5	82.5 $\pm$ 3.3	84.0 $\pm$ 0.9	84.1 $\pm$ 2.7	43.3 $\pm$ 2.0	54.7 $\pm$ 4.8
Con	Odd	Over	OFF	Spa	45.8 $\pm$ 3.1	79.1 $\pm$ 4.5	84.6 $\pm$ 1.5	85.5 $\pm$ 0.8	42.3 $\pm$ 4.2	53.1 $\pm$ 5.1
ViT	Std	Und	ON	Alp	83.0 $\pm$ 2.1	96.3 $\pm$ 1.0	66.7 $\pm$ 4.3	92.6 $\pm$ 2.4	47.1 $\pm$ 1.8	56.8 $\pm$ 1.5
ViT	Std	Und	ON	Spa	81.7 $\pm$ 3.2	95.4 $\pm$ 0.4	63.8 $\pm$ 0.9	92.4 $\pm$ 1.0	48.6 $\pm$ 2.1	58.4 $\pm$ 1.1
ViT	Std	Over	ON	Alp	84.8 $\pm$ 0.6	96.0 $\pm$ 0.2	64.7 $\pm$ 0.1	92.9 $\pm$ 1.4	46.8 $\pm$ 3.2	57.9 $\pm$ 1.2
ViT	Std	Over	ON	Spa	80.3 $\pm$ 1.5	95.6 $\pm$ 0.4	67.0 $\pm$ 0.7	92.8 $\pm$ 2.6	48.0 $\pm$ 0.8	58.3 $\pm$ 0.3
ViT	Std	Und	OFF	Alp	64.4 $\pm$ 2.5	97.0 $\pm$ 0.9	77.2 $\pm$ 2.6	93.4 $\pm$ 1.1	49.5 $\pm$ 2.6	58.9 $\pm$ 2.5
ViT	Std	Und	OFF	Spa	62.8 $\pm$ 4.8	96.5 $\pm$ 1.1	78.2 $\pm$ 2.0	93.8 $\pm$ 2.0	50.0 $\pm$ 1.0	60.9 $\pm$ 1.0
ViT	Std	Over	OFF	Alp	62.5 $\pm$ 4.1	97.0 $\pm$ 0.6	76.9 $\pm$ 1.3	94.0 $\pm$ 1.1	47.5 $\pm$ 2.1	60.2 $\pm$ 1.1
ViT	Std	Over	OFF	Spa	57.8 $\pm$ 2.2	97.7 $\pm$ 0.2	78.0 $\pm$ 1.4	94.4 $\pm$ 1.2	46.6 $\pm$ 0.8	61.4 $\pm$ 1.3

Table A1: Supplementary Table A1. ERPVR dataset F1-score from base LOSO and calibration phases averaged over random seeds and subject group (mean $\pm$ (max – min)/2). Results organized by configuration $\times$ subject group. 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.”



(a) ERPVR Subject 15's IGs averaged over(b) ERPVR Subject 10's IGs averaged over sick data from experiment run when balanced sick data from experiment run when balanced accuracy > 75%. accuracy > 75%.

Fig. B1: Supplementary Figure B1. ERPVR mixed label subjects that reached high balanced accuracy.

statistical significance for using Standards instead of Oddballs for classification, with Standards leading to increased balanced accuracy. And for electrode channel order, it seems there is no statistical significance for classification.

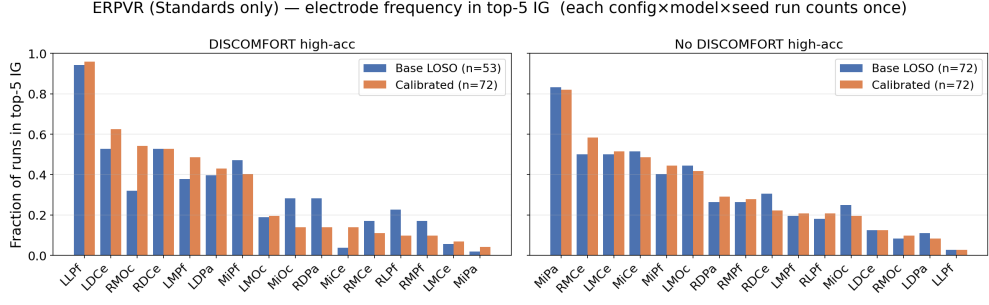
Transformation	n	Mean Δ (%)	Median Δ (%)	Wilcoxon
Outlier detection (OFF - ON)	29	+1.74	+2.84	W=159, p=0.213
Sampling (over - under)	29	-0.12	-0.17	W=190, p=0.565
Data type (Oddballs - Standards)	29	-2.02	-2.69	W=102, p=0.011
Channel order (spatial - alpha)	29	-0.22	-0.29	W=174, p=0.358
Calibration (calib - base LOSO)	29	+9.64	+3.71	W=54, p=0.000174

Table B2: Supplementary Table B2. Wilcoxon signed-rank test on balanced accuracy from ERPVR runs on all subjects (n=29). Each run in the paired test comparison aggregates per-subject balanced accuracy differences Δ , one averaged value per subject over its matched pairs. n is the number of contributing subjects.

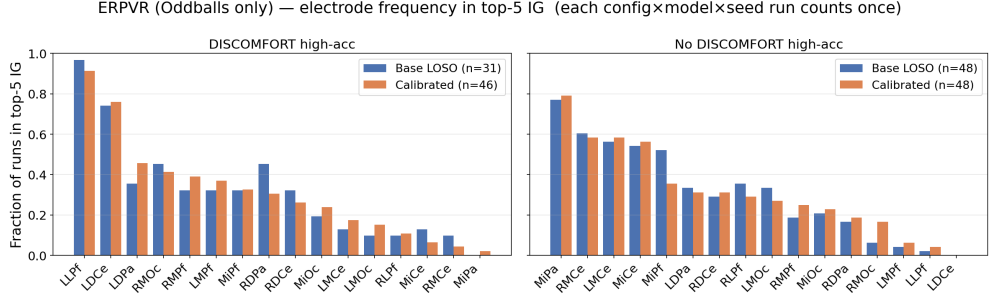
Appendix C Additional details on data analysis and augmentation

C.1 Individual trial-level outlier detection

With the strict outlier detection and removal method outlined in section 7.3, approximately one-third of the nonsick trials and one-third of the sick trials were identified as outliers in ERPVR. Exact numbers varied across random seeds depending on which subjects are left out as validation and test. An example breakdown of the outlier data removal by subject is shown for nonsick in Supplementary Figure C6 and for sick in Supplementary Figure C7. In this case, approximately one-third of nonsick



(a) ERPVR fraction of runs using Standards where electrode was in the top-5 integrated gradients for high-accuracy subjects.



(b) ERPVR fraction of runs using Oddballs where electrode was in the top-5 integrated gradients for high-accuracy subjects.

Fig. B2: Supplementary Figure B2. ERPVR fraction of runs where electrode was in the top-5 integrated gradients for high-accuracy subjects. n is number of experiments where the setting had subjects classified with high accuracy, discomfort subjects on the left and no discomfort right.

and sick data were removed from a total of 6,108 nonsick data trials and 2,808 sick data trials.

C.2 Subject-level outlier detection

Since there was high inter-subject variability in EEG data, we visualized how different each subjects' data was compared to others out of the 29 subjects in the ERPVR study. For subject outlier detection, we examined the plots from data analysis. Plots were made by averaging across the EEG data between sick and non-sick labels and between the 28 subjects vs. the one subject left out, providing a qualitative view of the data (see Figure C8). For subjects whose data visually looked significantly different from the average of the remaining 28 subjects, we checked the notes from the cybersickness study for a possible explanation. For example, after observing the variation when Subject 10 was left out (Figure C8), we discovered from the notes that they felt extremely sick to the point of nearly throwing up, whereas the other participants did not feel such an

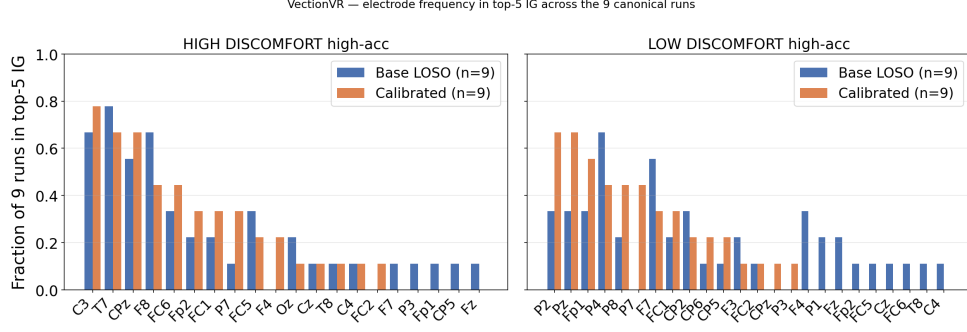


Fig. B3: Supplementary Figure B3. VectionVR fraction of runs using Standards where electrode was in the top-5 integrated gradients for high-accuracy subjects. n is number of experiments where the setting had subjects classified with high accuracy, discomfort subjects on the left and no discomfort on the right.

Transformation	Pairs	Group	n	Mean Δ (%)	Med Δ (%)	Wilcoxon
Calibration (Calib - base LOSO)	120	AD	6	+25.10	+18.06	W=0, p=0.031
Calibration (Calib - base LOSO)	120	mixed	8	+1.67	+1.72	W=3, p=0.039
Calibration (Calib - base LOSO)	120	ND	15	+7.71	+3.71	W=26, p=0.055
Outlier detection (OFF - ON)	60	AD	6	-23.24	-22.45	W=0, p=0.031
Outlier detection (OFF - ON)	60	mixed	8	+0.56	+0.15	W=15, p=0.742
Outlier detection (OFF - ON)	60	ND	15	+12.35	+10.85	W=0, p=6.10e-05
Sampling (over - under)	60	AD	6	-0.12	-0.95	W=6, p=0.438
Sampling (over - under)	60	mixed	8	-0.73	-0.78	W=14, p=0.641
Sampling (over - under)	60	ND	15	+0.20	+0.22	W=54, p=0.762
Data type (Oddballs - Standards)	48	AD	6	+1.12	+0.95	W=7, p=0.562
Data type (Oddballs - Standards)	48	mixed	8	-0.15	+0.55	W=15, p=0.742
Data type (Oddballs - Standards)	48	ND	15	-4.28	-4.33	W=6, p=8.54e-04
Channel order (spatial - alpha)	60	AD	6	-0.52	-0.08	W=7, p=0.562
Channel order (spatial - alpha)	60	mixed	8	+0.48	+0.76	W=13, p=0.547
Channel order (spatial - alpha)	60	ND	15	-0.48	-0.82	W=36, p=0.188

Table B3: Supplementary Table B3. Wilcoxon signed-rank tests by effect by configuration and subject group from all ERPVR runs over balanced accuracy. AD means “all discomfort” and ND means “no discomfort”.

intense level of sickness. Another method used was observing if there were subjects that all models performed poorly on. For example, all models in this work did not evaluate well on Subject 26 prior to calibration (Figure C9). As mentioned in the study notes, Subject 26 had very noisy data from jaw clench muscle contractions. Although there were individuals that deviated quite a bit from the average of the remaining subjects, we did not remove any from the data because these distribution shifts were representative of how different people can and will have EEG variability.

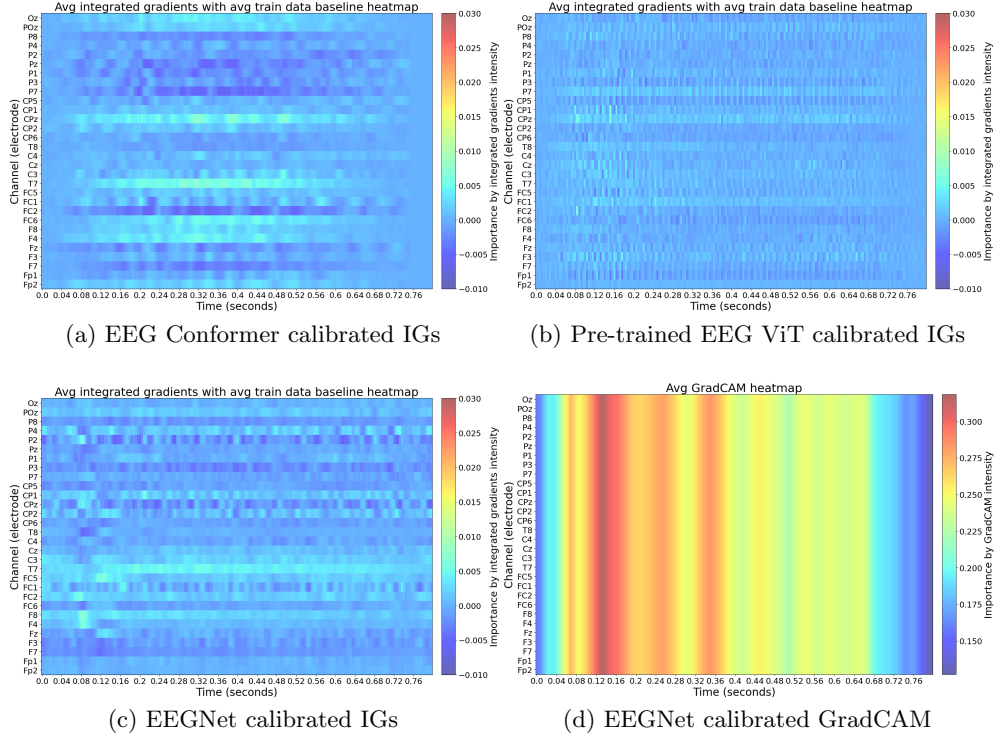


Fig. B4: Supplementary Figure B4. VectionVR Integrated Gradients (IGs) averaged over discomfort data from subjects with balanced accuracy > 75% (from seed 3). Warmer colors indicate more positive attribution.

C.3 Synthetic data for class imbalance

A common method for generating synthetic data is Synthetic Minority Over-sampling Technique (SMOTE) [2]. We explored several variants [3] using the smote-variants library [4]. Since SMOTE uses kNN to interpolate synthetic data, it does not work well for high dimensional and noisy data like EEG. In earlier test runs, we observed that neural networks like EEGNet performed better with random oversampling. Thus, SMOTE was not used for the deep learning pipeline. However we did find Safe-level SMOTE, which generates synthetic data by using “safe” samples identified by kNN, worked well for our data after it has been projected to PCA space to reduce dimensions. The principal components, both real and synthetic, were passed into traditional non-neural network models for classification.

Appendix D ML for Pattern Recognition

We considered both supervised and unsupervised learning methods to extract patterns from the ERPVR data. With supervised learning, both traditional and neural networks were explored. CNN and vision transformers were found to perform the best across the LOSO cross-validations, so we focused on those models for interpretation.

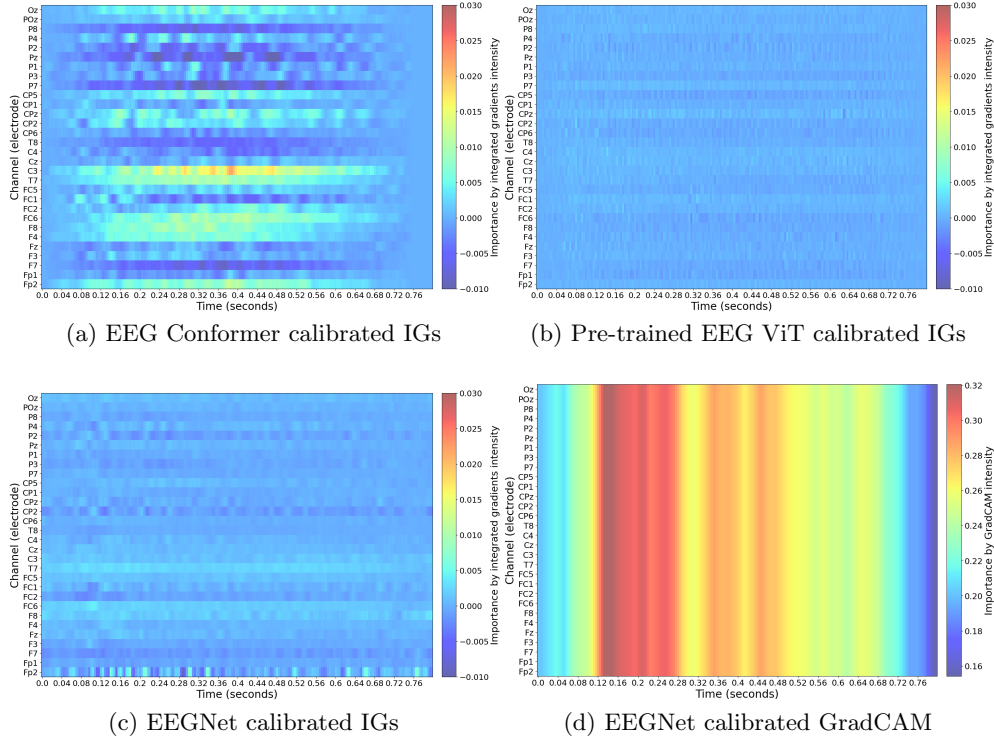


Fig. B5: Supplementary Figure B5. VectionVR Integrated Gradients (IGs) averaged over discomfort data from subjects with balanced accuracy $> 75\%$ (from seed 2). Warmer colors indicate more positive attribution.

D.1 Unsupervised learning

Clustering was run to see if any patterns could be identified from individual channel EEG without any labels in the ERPVR data. We used k-means clustering, where 6 to 7 clusters were found to be the ideal number based on silhouette score [5]. Silhouette score measures inter and intra-cluster distance for each data sample. Each cluster centroid is plotted in Figure D10. To visualize the high dimensional data, we employed T-distributed Stochastic Neighbor Embedding (t-SNE) [6] for two components, seen in Figure D11. From the right plot, we can observe that separation between sick and non-sick labeled data is difficult. When decomposing data in each cluster, each subject and electrode had some data in each cluster, indicating that the EEG data were difficult to classify across subjects and electrodes.

D.2 Supervised learning with traditional ML models

We tried to find underlying patterns present in the data using the binary sickness labels by comparing traditional machine learning model performance based on their inductive biases. Though not as powerful as neural networks in capturing complex patterns, traditional models have the advantage of requiring less train data and being more interpretable. We evaluated multiple traditional models with LOSO cross validation.

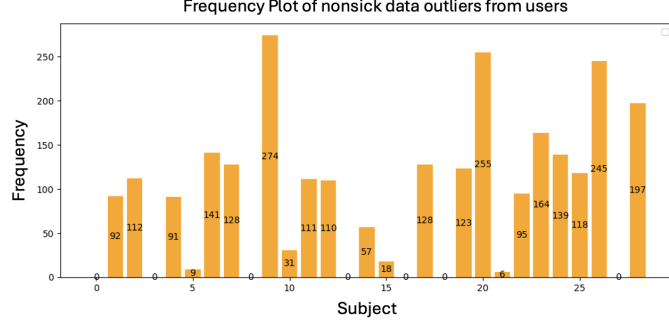


Fig. C6: Supplementary Figure C6. Example trial of nonsick data outlier removal frequency plot by subject with subject 0 left out as test and subject 27 as validation

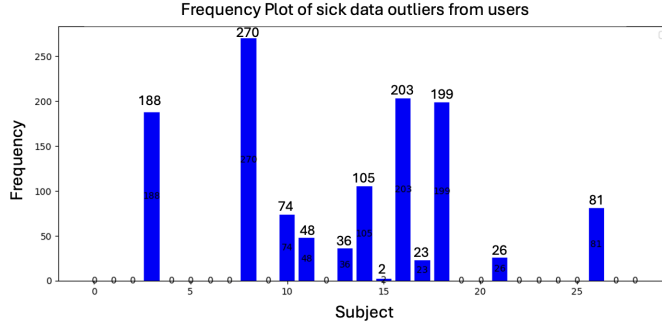


Fig. C7: Supplementary Figure C7. Example trial of sick data outlier removal frequency plot by subject with subject 0 left out as test and subject 27 as validation

The results averaged over all LOSO-folds are shown in Table D4. Scikit-learn [1] was used for implementation. The input data was a flattened array of number of electrodes (channels) $C = 16$ by time T of size $C * T$. Our pipeline removed outliers and normalized with z-score before running PCA for 48 components and safe-level SMOTE.

Appendix E Deep Learning Model Descriptions

EEGNet. EEGNet [7] is a two block CNN that has both spatial depthwise convolutions across the channels and temporal convolutions along the time dimension. In the first block is a temporal convolution with kernel size $(1, 64)$ and stride 1 along the time dimension, followed by a spatial convolution with kernel size $(C, 1)$ and stride 1 to learn a spatial filter across all electrodes, where C is the number of electrodes. The second block has a separable (depthwise) convolution followed by a pointwise convolution to separate learning how to summarize individual feature maps in time (the depthwise convolution) with how to optimally combine the feature maps (the pointwise convolution) [7]. Finally, a fully-connected layer outputs classification logits.

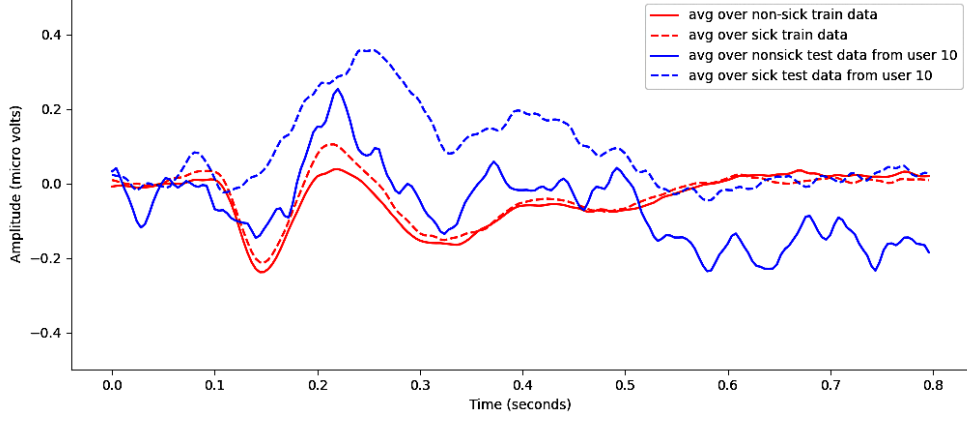


Fig. C8: Supplementary Figure C8. Leaving out Subject 10 from the remaining subjects.

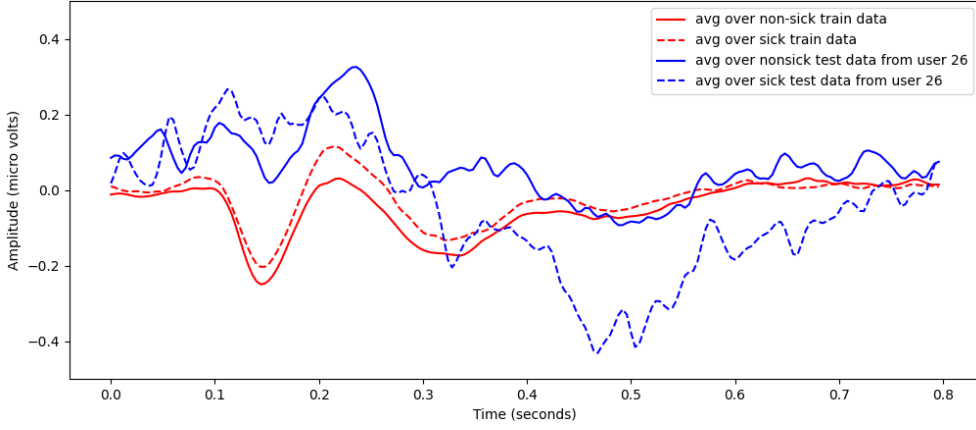


Fig. C9: Supplementary Figure C9. Leaving out Subject 26 from the remaining subjects

EEG Conformer. EEG Conformer [8] is a convolutional transformer with three parts: a convolution block to create feature maps and form tokens, a transformer to model relationships within the token sequence, and a classifier block. The convolution block convolves over the EEG signal similarly to EEGNet’s temporal and spatial convolution, with temporal kernel (1, 25) and stride 1, and spatial kernel (C , 1) with stride 1. The learned filters are pooled and rearranged into a sequence of tokens to be passed to the transformer, composed of four encoder blocks. Each encoder block employs multi-head self-attention to learn longer-term temporal features in the token sequence. The self-attention focuses on the relationship between tokens in sequence while multiple heads can represent and capture different kinds of relationships simultaneously. In this work, we use 8 heads and 40 hidden size. The classifier block is composed of three fully-connected layers, outputting classification logits.

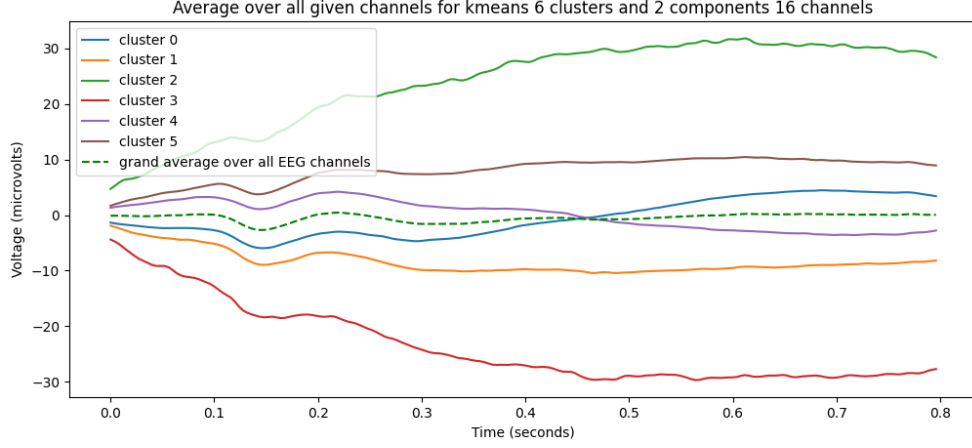


Fig. D10: Supplementary Figure D10. K-means centroids of 6 clusters and average over ERPVR data.

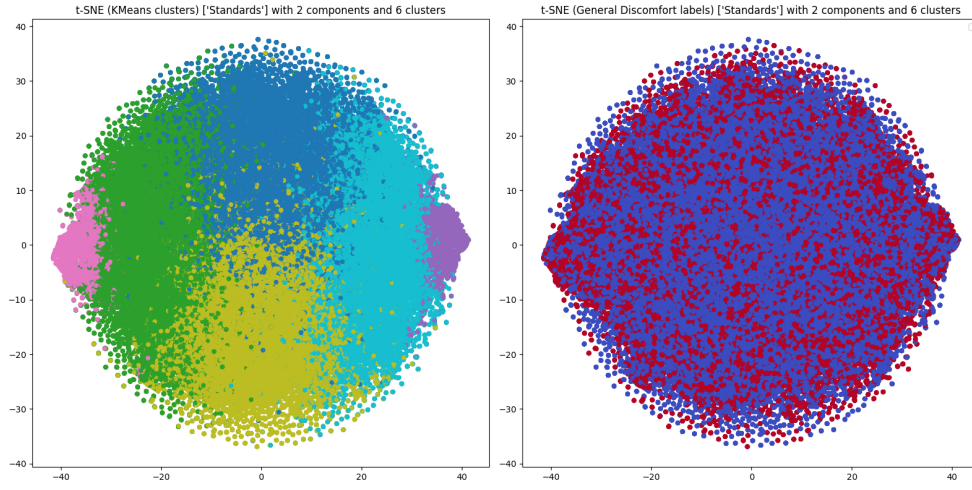


Fig. D11: Supplementary Figure D11. t-SNE for 2 components of ERPVR data. Left: data points from the same cluster have the same color. Right: data points are colored based on sick and non-sick label.

Pre-trained EEG ViT. Vision Transformers extract information from an image by decomposing it into a sequence of patch embeddings and learning relationships between the patches with the transformer blocks. We modified EEG ViT [9] for our task. This model has a patch embedding convolution block, a pre-trained transformer composed of 12 encoder blocks with 12 heads, and a classifier block. It was pre-trained on the ImageNet dataset so it learned to extract patterns from images. Since it was specifically trained on 224×224 size images, decomposed into patches of size 16×16 and has 196 patch tokens in the patch embedding sequence along with the class token

kNN	AdaBoost	RF	LDA	QDA	B NB	G NB	M NB	SVM	GP	Relu MLP
0.5823	0.6510	0.5566	0.5031	0.5628	0.5066	0.5473	0.4785	0.5983	0.5612	0.6129

Table D4: Supplementary Table D4. ERPVR average balanced accuracy for traditional ML models across all LOSO cross validations with outlier removal, normalization, PCA, and safe-level SMOTE using the SSQ general discomfort symptom intensity as labels. Results are reported for k=2, 10-tree AdaBoost and random forest (RF) with log loss split criterion, Bernoulli Naive Bayes (B NB), Gaussian Naive Bayes (G NB), Multinomial Naive Bayes (M NB), Singular Value Decomposition LDA, quadratic discriminant analysis (QDA), polynomial kernel SVM, and Radial-Basis Function kernel Gaussian Process (GP).

for 197 tokens with 768 hidden size [10], modifications were made to match those dimensions given our EEG input data. Our modified model for ERPVR starts with a temporal convolution of (1, 35) with stride 3 and creates patch embeddings with kernel size (4, 1) to output 224 tokens of 4×1 size patches and the class token for 225 total tokens. For VectionVR, the patch embeddings are created with kernel size (6, 1) to output 280 tokens of 6×1 size patches and the class token for 281 total tokens. The token sequence is passed to the pre-trained ViT backbone and two-layer fully-connected classifier block, outputting classification logits.

Appendix F Model Interpretability

Integrated gradients. To visualize which input features the neural network models found helpful (or not) for the model’s classification of the target class label, integrated gradients [11] were extracted using the Captum library [12]. Integrated gradients (IG) is an axiomatic model interpretability algorithm that assigns an importance score to each input feature by approximating the integral of gradients of the model’s output with respect to the inputs along a straight line path from a chosen baseline to the input [12] by accumulating gradients along the path. IG satisfies the completeness axiom that states attributions add up to the difference between the output of the model F at the input x and the baseline x^0 [11]. It provides a signed value for each input feature, positive attribution for helping the model classify to target label and negative for the reverse. In our method, the baseline was the grand average over all train data (close to zero due to normalization).

GradCAM. Gradient-weighted class activation maps (GradCAM) were extracted with the Pytorch CAM methods library [13]. GradCAM shows which activations in the feature map had more focus for classification of a target label by taking the gradient with respect to the chosen convolution layer’s feature map.

Interpretability for each model. 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 sickness. 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, 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.

Appendix G Experiment Setup Details

For the main training phase, we used an Adam optimizer with learning rate and weight decay of 1×10^{-4} . A multi-step learning rate scheduler was used for EEGNet and EEG Conformer, with milestones at 10,000 and 20,000 steps. For the pre-trained EEG ViT, cosine scheduler with a warmup of 1,000 steps was used, as done by the original model [10]. The batch size was 64. Dropout was 0.25 for both EEGNet and EEG Conformer while pre-trained EEG ViT was 0.1. Cross entropy loss was used (focal loss was tested with the class label imbalance but with random oversampling to balance each batch, cross entropy was found to perform better). For EEGNet and EEG Conformer, the models were trained from scratch. For the pre-trained EEG ViT, fine-tuning was run where the 12 encoder blocks were frozen while the initial convolution and batchnorm, patch embedding, and final classifier layers were unfrozen. For VectionVR runs, pre-trained EEG ViT also had positional embeddings unfrozen. For the transformer models, gradient clipping was applied to stabilize training. Balanced accuracy was the metric used in validation to determine early stopping point. For each stimulus in ERPVR, 800 ms from 16 EEG electrodes (channels) were used as input data to the models, a matrix of shape $C = 16$ by $T = 800$ ms. For VectionVR, it is 800 ms from 30 EEG electrodes (channels) used as input data to the models, a matrix of shape $C = 30$ by $T = 800$ ms.

Next was calibration, where the best validation checkpoint for each model for each subject left out was fine-tuned on the smaller calibration dataset with a constant learning rate of 1×10^{-5} . Cross entropy loss was weighted per sample (eq. 3), where data from the test subject had double the weight of the data from the others for subjects who only had one label. A custom weighted balanced accuracy (eq. 5) was used to determine early stopping points.

Some EEGNet and EEG Conformer experiments were run with an Nvidia RTX 4060 GPU, but most were run on an Nvidia RTX 4090. The pre-trained EEG ViT experiments were run on an Nvidia RTX 4090. All are run on Ubuntu Linux computers.

Appendix H Additional plots

We include additional integrated gradient plots from different random seed experiment trials in Figs. H13 and H14 for ERPVR. The plots for the trial with channel order rearranged is in Figure H12. LLPf remains the electrode channel with highest attribution across the models for classifying sickness in ERPVR.

H.1 Using different SSQ symptom as label

To verify our method works for other kinds of patterns, “Difficulty focusing,” a different SSQ symptom, was used to create binary classification labels for ERPVR in Figure H15. The top attribution channels are RDCe and MiPf this time, highlighting different electrodes are useful for detecting “Difficulty focusing”.

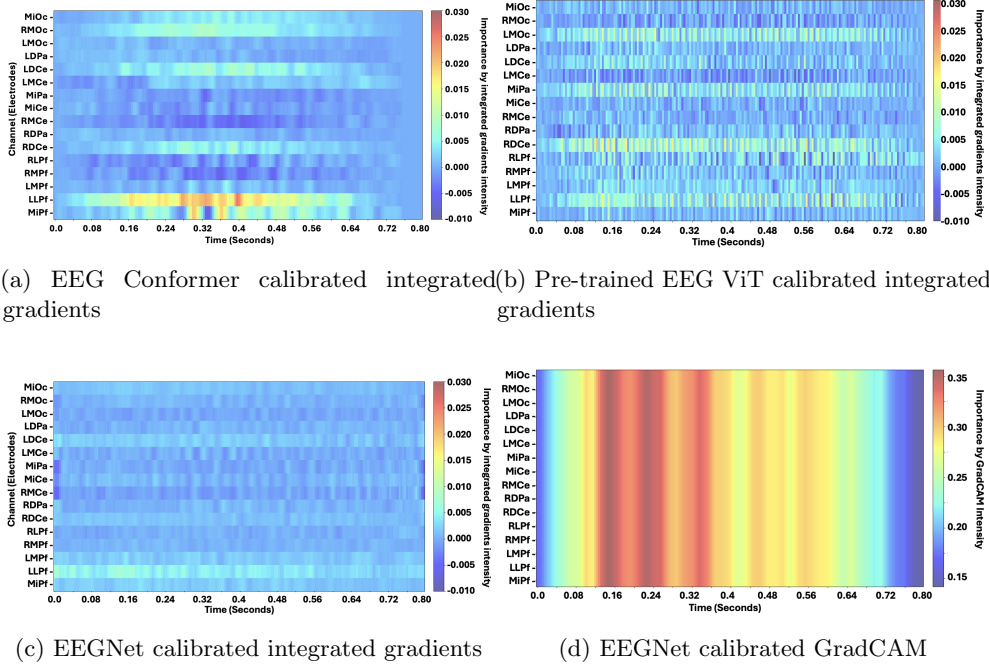


Fig. H12: Supplementary Figure H12. Integrated gradients for calibrated models averaged over sick data from balanced accuracy $> 75\%$ subjects and GradCAM for EEGNet trained with different channel order. Warmer colors indicate more positive attribution.

Appendix I ERP Waveform Analysis by Electrode Groups

After identifying electrodes of interest from the ERPVR dataset and VectionVR dataset, we plot waveforms by averaging over the identified top 8 consistent electrodes for discomfort vs. no discomfort. For ERPVR no discomfort, the top 8 electrodes that contributed to classification were: MiPa, RMCE, LMCE, MiCe, LMOc, RMPf, MiPf, RDPa. For ERPVR discomfort, the top 8 electrodes that contributed to classification were: LLPf, LDCe, RDCe, RMOc, LDPa, LMPf, MiPf, MiOc. Both are plotted in Supplementary Figure I16 for Standards waveforms. Oddballs are at Supplementary Figure I17. For VectionVR for high discomfort, we have: C3, T7, CPz, F8, Fp2, FC1, Cz, Oz and for low discomfort, we have: P2, Fp1, Pz, F7, P4, P8, CP6, F3. The VectionVR averaged waveforms are in Supplementary Figure I18.

Appendix J ERPVR using SSQ Unweighted Symptom Sum as Label

To evaluate another labeling method instead of taking a single symptom of SSQ, we tried using the SSQ unweighted symptom sum (USS) as a label, where it is defined

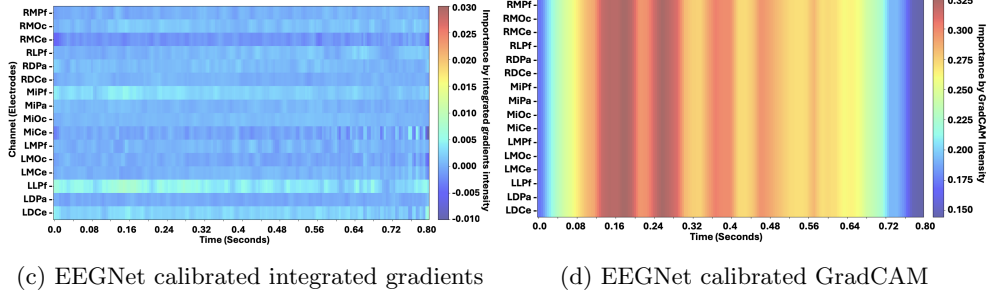
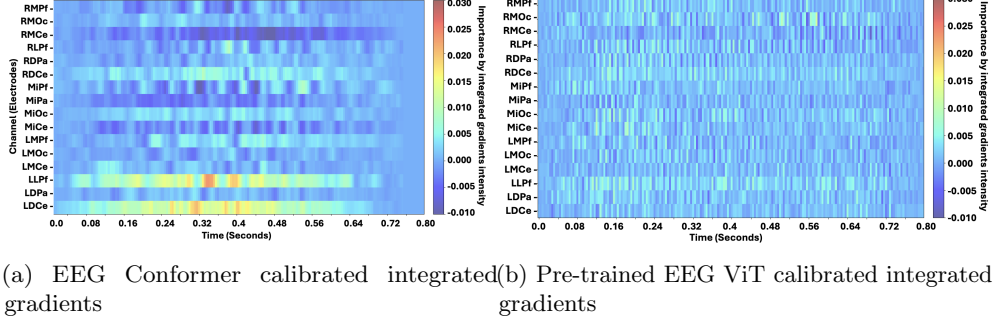


Fig. H13: Supplementary Figure H13. Integrated gradients for calibrated models averaged over sick data from balanced accuracy $> 75\%$ subjects and GradCAM for EEGNet with random seed 3. Warmer colors indicate more positive attribution.

as the sum across all 16 SSQ symptoms. Binary labels are created where $USS > 0$ is sick and $USS = 0$ is nonsick. With this labeling, the subjects split as so: 12 pure-sick, 12 mixed labels, and 5 pure-nonsick. We tested this labeling method with EEGNet and Standards data. Looking at subject 15 who was mixed label under the general discomfort labeling scheme, they also were mixed label under the USS labeling. So as a subject who has mixed labels in both labeling formats, it was interesting to look at how the model performs with the different labeling. However, subject 15's balanced accuracy after calibration dropped greatly (38% from 44.43%) when using the total score to set the binary classification labels, compared to the general discomfort labels (81%). This may be because when using the SSQ unweighted symptom sum, all symptoms are mixed together to compute the sum. With all SSQ symptoms mixed together, it may be harder for the models to extract specific patterns distinguishing between the sick and nonsick labels defined by the total score. Also, the very low balanced accuracy for subject 15 at base LOSO indicates that the training pool shift from the USS labels and the patterns the model learned from this shifted train set is misaligned to subject 15 such that even calibration is not able to recover the model performance on the subject.

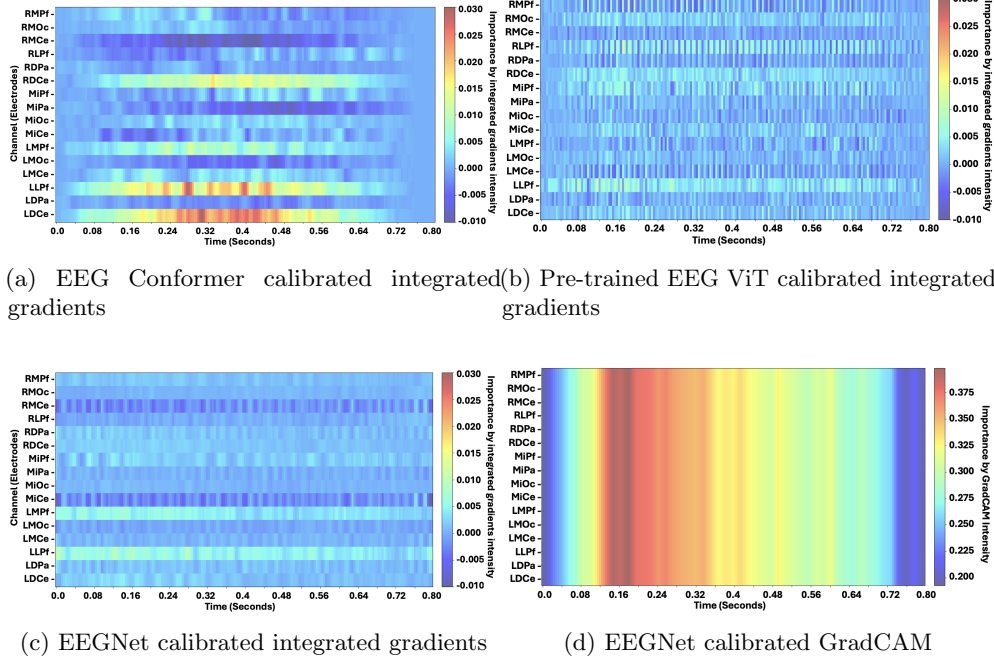


Fig. H14: Supplementary Figure H14. Integrated gradients for calibrated models averaged over sick data from balanced accuracy > 75% subjects and GradCAM for EEGNet with random seed 2. Warmer colors indicate more positive attribution.

Appendix K Measuring Cybersickness with the P3b

The P3b ERP in EEG (also called the P3 or P300) is one of the most well-studied ERP components [14]. Though there are several theories as to what specific cognitive process the P3b reflects, across decades of research and hundreds of studies, the amplitude of the P3b has consistently been shown to be influenced by task difficulty, cognitive load, and attention [15]. As such, it has been used in myriad studies to investigate attentional allocation [14–16].

Attention can be conceptualized as a cognitive resource for prioritizing processing to achieve a goal [17]. Over the years, research on attention has shown that attentional changes are more apparent in brain or behavior measures when there is greater competition for processing resources [14, 15, 17]. Thus, a common method to study attention is to use a dual-task paradigm [18]. In dual-task paradigms, participants are given two tasks that must be performed together, sometimes in different sensory modalities. When there is a primary visual task, attentional changes can be measured with the P3b from an auditory secondary task, like the oddball paradigm [14]. Auditory oddball paradigms involve having participants listen to a series of beeps. Infrequently and randomly, a beep will sound different from the other beeps. Participants are asked to pay attention to these series of beeps and respond in some way, for example by

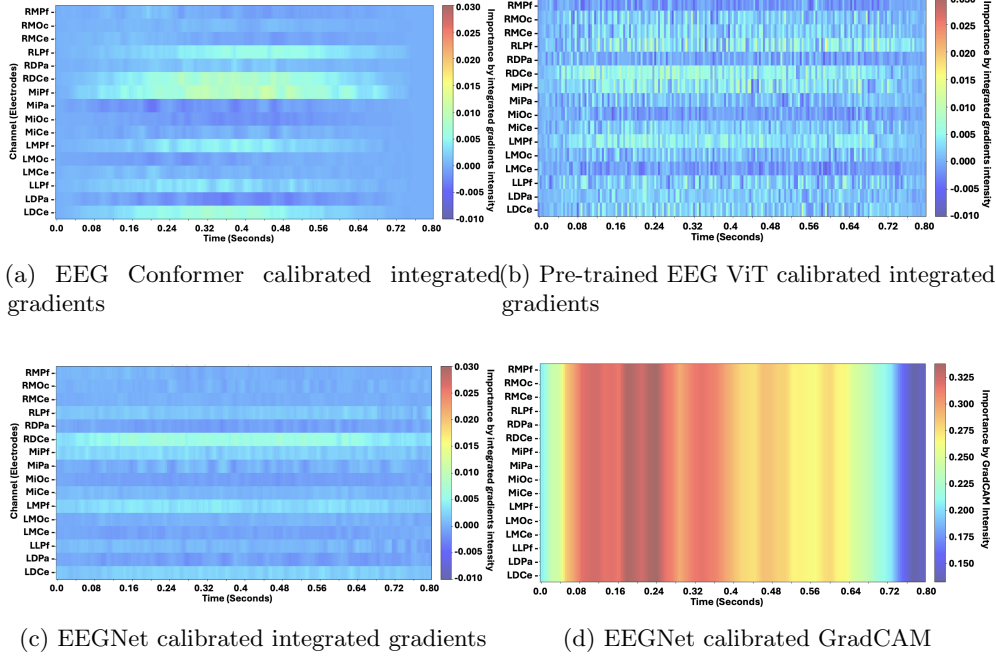
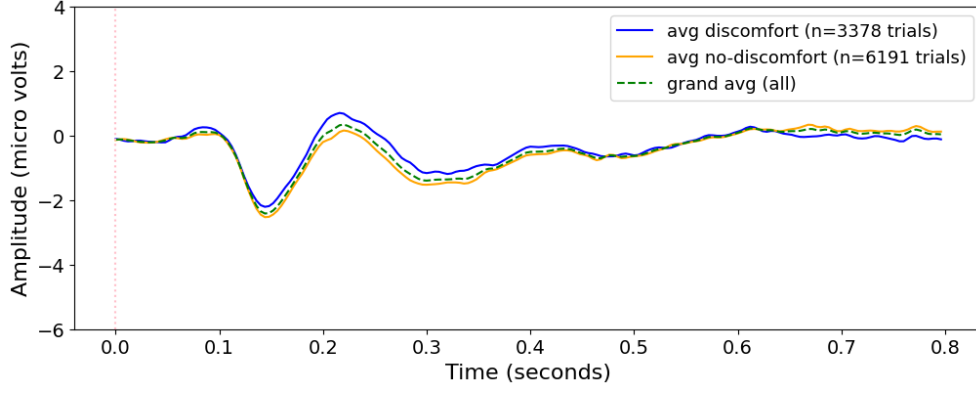


Fig. H15: Supplementary Figure H15. Integrated gradients for calibrated models averaged over sick data from balanced accuracy > 75% subjects and GradCAM for EEGNet for “Difficulty focusing” labels. Warmer colors indicate more positive attribution.

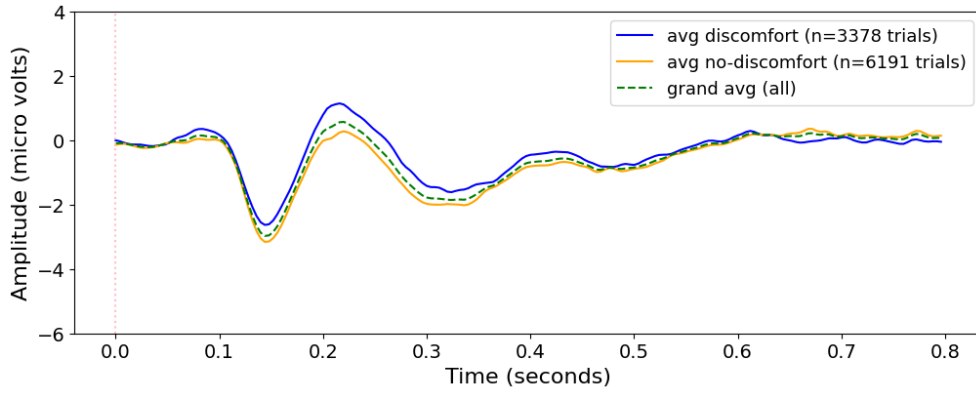
keeping a mental count of how many different sounding beeps they heard [16]. This simple method to produce the P3b and measure attention is reliable and sensitive enough that it has been used in at least 35 studies with more than 1,500 participants to investigate attentional differences in neurological disorders [16].

Utilizing this robust and empirically validated attentional measure, we [19] showed that cybersickness impacts attention. Using P3b amplitudes to an auditory secondary task during a primary visual task in virtual reality, we demonstrated that amplitudes of the P3b were directly associated with the amount of cybersickness that participants experienced, with reduced P3b amplitudes during periods of greater sickness. Consequently, the ERPVR EEG data consists of hundreds of periods (trials) of brain activity generated in response to the exact same stimulus (the beep) for each person. Embedded within these trials are very similar patterns of auditory cognitive processing which are subtly modified by the amount of discomfort experienced. Although the brain activity pattern from the auditory stimulus repeats hundreds of times (providing a rich source of data to be parsed algorithmically), the visual stimuli from the VR environment were different during each trial. Thus, the randomness of the visual cognitive processing essentially becomes minor background noise during



A(8) = LLPf, LDCE, RDCE, RM0c, LDPa, LMPf, MiPf, Mi0c

(a) ERPVR ERP Standards waveforms averaged by top eight contributing electrodes for discomfort classification, per label.

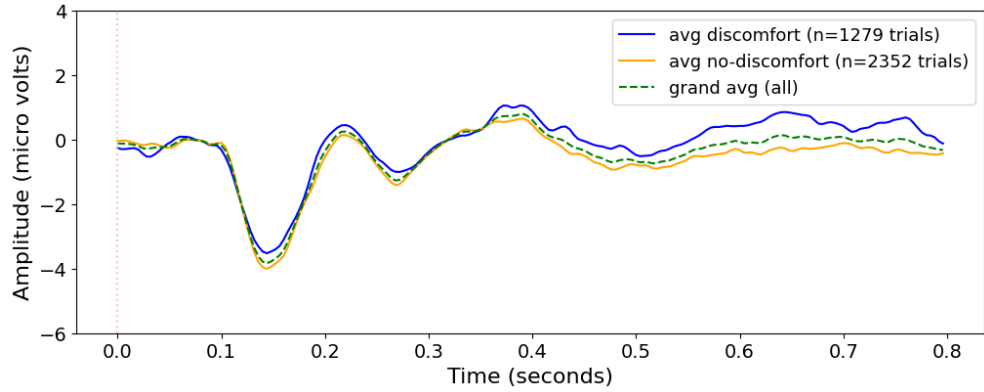


A(8) = MiPa, RMCE, LMCE, MiCE, MiPf, LM0c, RDPa, RMPf

(b) ERPVR ERP Standards waveforms averaged by top eight contributing electrodes for no discomfort classification, per label.

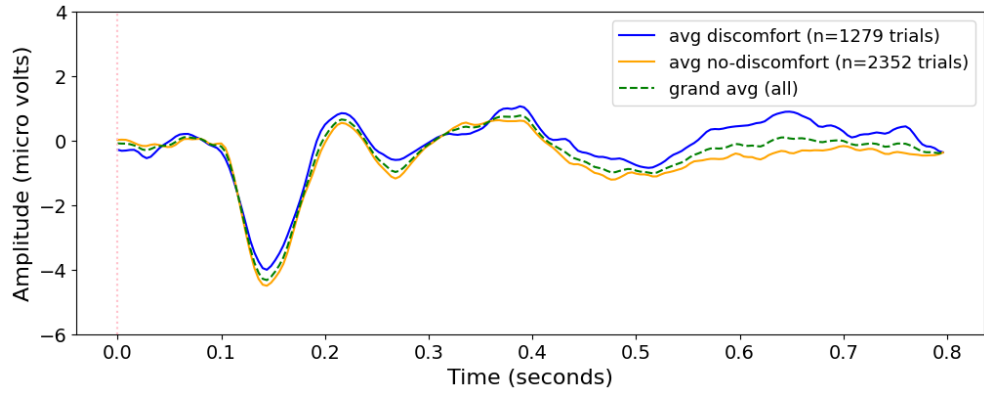
Fig. I16: Supplementary Figure I16. ERPVR Standards ERP waveforms averaged by identified electrode subset that contributed positively for model classification for each label. Vertical pink dotted line indicates when the audio stimulus was played, around 0.0 seconds.

each largely similar window of auditory stimulus processing. As a result, this ERP data controls for a visual confound that usually exists in EEG cybersickness data.



A(8) = LLPf, LDCE, LMPf, RMOC, LDPa, RDCE, MiPf, RDPa

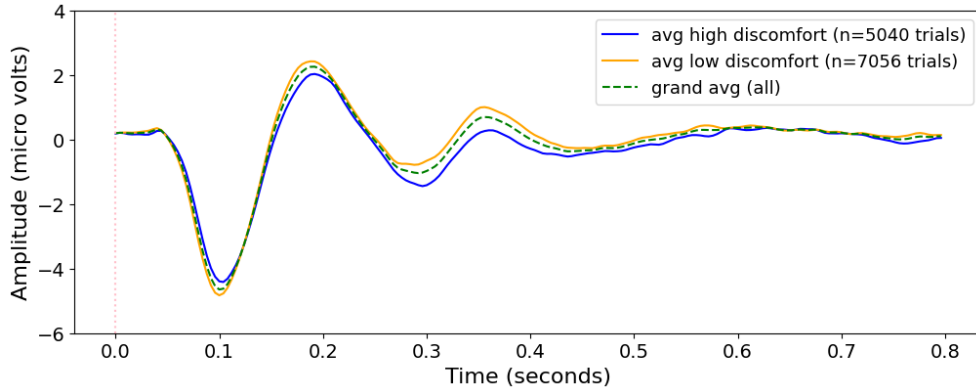
(a) ERPVR ERP Oddballs waveforms averaged by top eight contributing electrodes for discomfort classification, per label.



A(8) = MiPa, LMCE, RMCE, MiCE, MiPf, LMOc, RLPf, RMPf

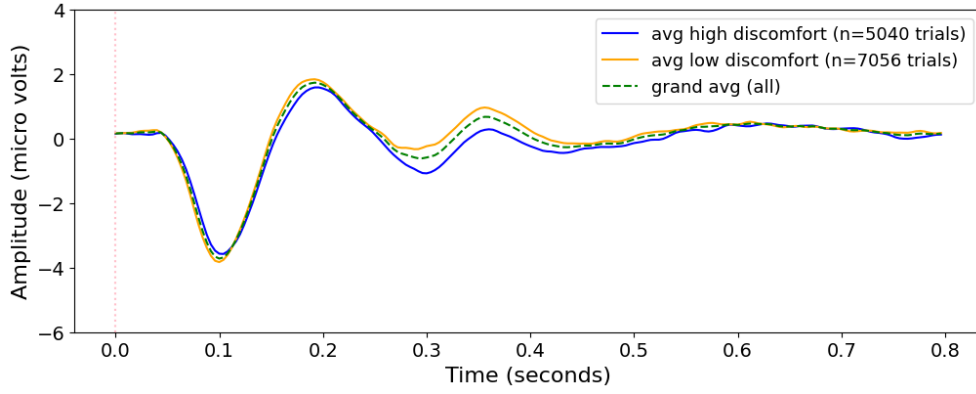
(b) ERPVR ERP Oddballs waveforms averaged by top eight contributing electrodes for no discomfort classification, per label.

Fig. I17: Supplementary Figure I17. ERPVR Oddballs ERP waveforms averaged by identified electrode subset that contributed positively for model classification for each label. Vertical pink dotted line indicates when the audio stimulus was played, around 0.0 seconds.



A(8) = C3, T7, CPz, F8, Fp2, FC1, Cz, Oz

(a) VectionVR ERP Standards waveforms averaged by top eight contributing electrodes for high discomfort classification, per label.



A(8) = P2, Fp1, Pz, F7, P4, P8, CP6, F3

(b) VectionVR ERP Standards waveforms averaged by top eight contributing electrodes for low discomfort classification, per label.

Fig. I18: Supplementary Figure I18. VectionVR Standards ERP waveforms averaged by identified electrode subset that contributed positively for model classification for each label. Vertical pink dotted line indicates when the audio stimulus was played, around 0.0 seconds.

References

- [1] 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)
- [2] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002) <https://doi.org/10.1613/jair.953>
- [3] Kovács, G.: An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing* **83**, 105662 (2019) <https://doi.org/10.1016/j.asoc.2019.105662> . (IF-2019=4.873)
- [4] Kovács, G.: smote-variants: a python implementation of 85 minority oversampling techniques. *Neurocomputing* **366**, 352–354 (2019) <https://doi.org/10.1016/j.neucom.2019.06.100>
- [5] Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987) [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- [6] Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of Machine Learning Research* **9**(86), 2579–2605 (2008)
- [7] 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>
- [8] 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>
- [9] Yang, R., Modesitt, E.: Vit2eeg: leveraging hybrid pretrained vision transformers for eeg data. *arXiv preprint arXiv:2308.00454* (2023)
- [10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021). <https://arxiv.org/abs/2010.11929>
- [11] 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>

- [12] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for PyTorch (2020). <https://arxiv.org/abs/2009.07896>
- [13] Gildenblat, J., contributors: PyTorch library for CAM methods. GitHub (2021)
- [14] 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>
- [15] 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>
- [16] 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>
- [17] Nobre, A.C.K., Kastner, S.: *The Oxford Handbook of Attention*. Oxford University Press, (2014). <https://doi.org/10.1093/oxfordhb/9780199675111.013.040> . <https://academic.oup.com/edited-volume/41256>
- [18] Kasper, R.W., Cecotti, H., Touryan, J., Eckstein, M.P., Giesbrecht, B.: Isolating the neural mechanisms of interference during continuous multisensory dual-task performance. *Journal of Cognitive Neuroscience* **26**(3), 476–489 (2014) <https://doi.org/10.1162/jocn.a.00480>
- [19] 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>