Integrated gradients for enhanced interpretation of P3b-ERP classifiers trained with EEG-superlets in traditional and virtual environments

Vladimir Marochko[1,*,†], Jacek Rogala[2,†] and Luca Longo[1,†]

[1] *Artificial Intelligence and Cognitive load Research Lab, Technological University Dublin, Dublin, Republic of Ireland.*
[2] *Faculty of Artes Liberales, University of Warsaw, Warsaw, Poland*

Abstract

P3b event-related potentials are neural responses occurring around 300-500 ms after the presentation of a task-relevant target stimulus, usually evoked with the active visual oddball paradigm. Usually performed in laboratories, it can be moved to virtual environments, especially for subjects who cannot reach the physical experimental setting. However, the introduction of a virtual reality headset introduces significant portions of artefacts, since it is overlapped on top of the EEG headset. This can hamper the analysis of the p3b ERP and thus limit the understanding of attentional allocation, working memory operations, and decision-making processes in the brain of a subject. The research aims to compare the quality of the EEG data collected in virtual environment against that collected in the traditional environment. After segmenting such data into epochs around each stimulus, superlets are computed for each, specific time-frequency representations, and trained with a convolutional neural network for the binary discrimination of segments collected in virtual and traditional settings. Explainable Artificial Intelligence, namely the Integrated Gradients method, is employed to facilitate the understanding of such discrimination. An equivalence two one-sided tests (TOST) is performed to verify if both the types of input EEG segments are statistically similar. This research contribute to the body of knowledge by extending the application of the Integrated Gradients XAI method in a problem within Neuroscience.

Keywords

Event-related potentials, Deep learning, Convolutional neural networks, Explainable Artificial Intelligence, Integrated Gradients, P3b, Oddball paradigm, time-frequency super-resolution, Superlets.

1. Introduction

Stimuli and response-related fluctuations in human brain electric activity, known as event-related potentials (ERP) [1], can provide meaningful information on the various human cognitive processes. This makes them a valuable tool for a broad range of applications in neuro and cognitive sciences research, diagnostics, including the support of humans in sensory-motor functions [2]. Despite the advantages in extracting ERP from EEG data, performing experiments in naturalistic contexts, in contrast to lab-based controlled settings [3], is not a trivial task. VR

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

[*] Corresponding author.

[†] These authors contributed equally.

vladimir.marochko@tudublin.ie (V. Marochko); Jacek.Rogala@fuw.edu.pl (J. Rogala); luca.longo@tudublin.ie (L. Longo)

0000-0001-9766-6420 (V. Marochko); 0000-0001-5298-920X (J. Rogala); 0000-0002-2718-5426 (L. Longo)

can support the creation of realistic scenarios and allow the manipulation of variables with precision, supporting ERP research in ecologically valid environments [4]. However, collecting EEG data in a VR environment is rather challenging because a VR headset influences the electrodes of an EEG cap. Therefore, an open problem is demonstrating the equivalence of ERP extraction in VR settings compared to traditional, lab-based settings [5]. This research is devoted to demonstrating such equivalence by employing deep learning for learning high-level representations from time-frequency representations and Explainable AI (XAI), namely the use of the Integrated Gradient method, for supporting the understanding of the difference between VR-collected versus lab-collected EEG data.

The remainder of this article is structured as follows. Section 2 briefly describes related work on ERP, its analysis in virtual reality settings, and the XAI methods employed in the field. Section 3 is focused on the design of an empirical work to demonstrate their equivalence. Section 4 presents the experimental findings, followed by a critical discussion. Eventually, it summarises the contribution to the body of knowledge and highlights future work.

2. Related Work

This section focuses on Event-Related Potential analysis and research on its application in Virtual Reality (VR). A trend is to apply Head Mounted Displays (HMD) to realise virtual reality, which allows the creation of a 360-degree environment (Fig. 1 right). VR-based environments are accessible due to a lack of mobility requirements, and are also ecological because of the higher level of control by the creator [3]. Unfortunately, despite these advantages, recording EEG data in such settings can increase the contamination of the EEG signal with artefacts, caused by the movement of the VR headset, usually placed on top of an EEG cap [4]. It was reported that the participant's eye, body and head movements in a VR-based experiment are more frequent and can lead to stronger movement artefacts than those in a traditional environment [6, 7]. A study reported a special modification of an EEG cap to decrease physical strain and minimise the noise generated by vertical movements. The modification involved cutting the holes in the VR-HMD straps where the straps cover the electrodes. For example, the study also proposed removing the batteries from the head and attaching them to the participant's belt [7]. After pre-processing and cleaning EEG data, the usual way of extracting an ERP component is to average the waveform of specific channels over epochs and then analyse its amplitude and latency from the onset of stimulus presentation. To support this, an informative way is to translate an EEG timeserie into a time-frequency representation, which can be subsequently averaged [8] Unfortunately, ERPS emerge only after such averaging procedures across trials and subjects because of the low EEG signal-to-noise ratio. This hampers the ability to highlight EEG neural signatures over trials. To tackle this issue, novel ERP analysis approaches based on deep learning have been devised. For example, a deep learning-based workflow, based on convolutional neural networks (CNN), has been presented in [9]. Similarly, CNNS were used to discriminate adult Attention Deficit Hyperactivity Disorder (ADHD) from healthy individuals using spectral information of EEG responses to stimuli [10]. The problem with deep learning models is that their internal structure is very opaque and hidden from human perception. For example, studies have already attempted to explain CNN models to classify ERPS by evaluating the effect of different parts of the input signal on the output [9]. These pieces of work have been developed within the larger field of Explainable Artificial Intelligence (XAI), which allowed scholars to propose generally applicable methods. These include Grad-cam [11] for the generation of visual explanations of black-box trained neural networks, or rule-based systems [12], or systematic interpretability of the latent space of neural networks [13]. Considering the abundance of gradient-based learning, a popular XAI method is the Integrated Gradients (IG)[14]. In supervised classification tasks, integrated gradients visualise the influence of different input data features, usually pixels in an image, on the neural network's final prediction. This method has recently been applied in ERP analysis to highlight the difference between ERP of subjects with alcoholism versus control subjects, or target vs non-target events, by employing CNNs [15]. Leveraging the strength of deep learning to learn EEG signatures over trials, and explainable AI techniques to allow the understanding of such signatures, this study aims to demonstrate the equivalence of the ERPS collected in a VR setting versus those in a traditional lab-based setting.

3. Experimental design and methods

Following the open problem of demonstrating the equivalence between the ERPS collected in a VR setting versus a traditional, lab-based setting, this section introduces a comparative empirical work. The component chosen for investigation is the P300 component, one of the most commonly researched in the ERP field due to its high clearance, familiar and well-polished protocol [16]. It is an endogenous component caused by the brain response associated with the oddball task, with two different groups of stimuli, usually visual. In such tasks, the participant has to define whether the stimulus belongs to the smaller target group or the larger non-target one. P3b scalp distribution is defined as the amplitude change over the midline electrodes, namely over the Fz, Cz, Pz channels according to the 10/20 system (Figure 1, centre). This component has a sharp positive voltage spike of 300 ms after presenting the stimulus [17].

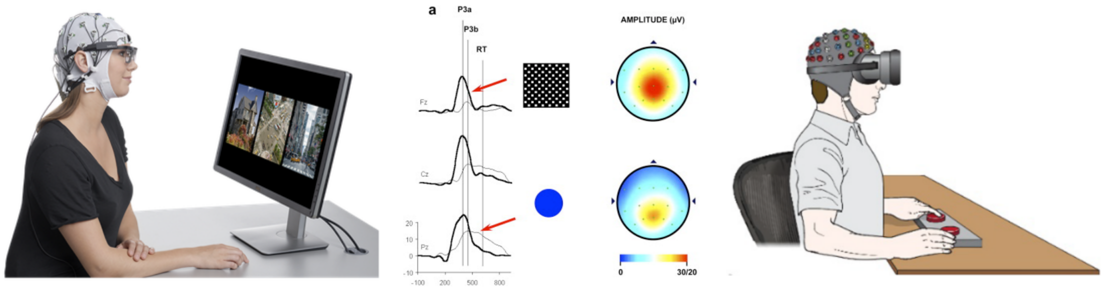

Figure 1: A traditional lab-based setting for Event-Related Potential (ERP) (left), a typical P3B waveform and a Virtual Reality-based setting

The experimental task is a classic Active Visual Oddball paradigm used for the P3b component extraction as demonstrated in the ERP-core research [17]. A participant is presented with a black screen where the letters A, B, C, D and E are shown in blocks. Each block has one target letter, and the others are non-target. In each block, 8 target and 32 nontarget letters are displayed. Letters are shown for 200 ms, with a random window of 1200 ± 100 ms between, when just a black screen with the white dot at the centre is shown. When presented with stimuli (letter), the participant should press the right trigger in VR or the 'c' letter on the keyboard key in the traditional environment for the target letter or the left trigger (in VR) or the 'z' letter on the keyboard (traditional) for the non-target. During the experiment, the 24-channel MbrainTrain Mobi EEG device records a participant's EEG waves and event ticks—time markers showing where the event has happened. The Oculus Quest 2 HMD headset creates the virtual reality setting. The Unity engine implements both traditional and VR environments. The OpenXR package assists in VR environments, and the LabStreamingLayer package sends event ticks to the MbrainTrain software. An Arduino-based system with a light sensor sending a signal to a LabStreamingLayer measures screen latency—the time lag between the event (letter shown) and the moment it appears on the computer or the VR HMD screen. The recorded latency is 40 ms for the screen and 70 ms for the HMD. Twenty-two healthy adult participants aged 20 to 50, both male and female, without a history of seizures or brain trauma, not receiving EEG-altering medication and with normal or corrected to normal eyesight, took part in the experiment. Half of the participants were asked to use the VR environment first, and the traditional one subsequently. In contrast, the other half did the opposite. This strategy meant eliminating the possible learning effects on the resulting ERPs.

The collected data was divided into four groups according to the independent feature: Target-VR, Target-Screen, Non-Target-VR, and Non-Target-Screen. The following pre-processing steps are executed: raw signals were high-pass filtered to 0.1 Hz, low-pass to 30 Hz and re-referenced to the average. Then the FASTER automated preprocessing pipeline was applied, using Fp1 and Fp2 instead of EOG channels [18]. Noisy channels were removed, the raw signal was separated into epochs (-200 ms, 800 ms to the stimuli), and too noisy epochs were removed. The ICA was then run, and bad channels were rejected and interpolated, always using the FASTER methodology. After such pre-processing, 18 participants were left with an average of 5% epochs excluded. Only data from the Pz electrode was taken for further investigation, as this is where the P3b signal is the strongest [17].

A CNN network (figure 2) was devised to train superlets, and it has been inspired by the EEGnet structure proposed in [19] with a changed layer to fit the scalogram input. L2- regularisation parameter of 0.000001, dropout rate of 0.25, depth multiplier 4, 64 temporal and 256 pointwise filters. Superlets are specific time-frequency representations of univariate EEG data that tackle the Heisenberg–Gabor uncertainty principle, which states that finite oscillation transients are difficult to localise simultaneously in time and frequency [20]. Superlets use sets of wavelets with increasingly constrained bandwidth, combined geometrically to preserve the good temporal resolution of single wavelets and gain frequency resolution in the upper bands. Superlets were generated for the interval 0-600 ms post-stimulus associated with target and non-target stimuli. After a few attempts, the 30 Hz bandwidth was separated into 60 bands, the cycle parameter was set to 3, and the order parameters from 3 to 20 were chosen to combine high transformation speed and good temporal and frequency resolution. The collected data was randomly split into training and testing sets with a ratio of 80/20, stratified by the independent feature (target or non-target), and the validation set was extracted from the former with an 85/15 split. The superlets in both the training and validation sets were augmented by adding a

Gaussian noise with an amplitude of 0.02%, essentially doubling their cardinality. Two models were trained, one for the data collected in the VR environment and one for the data collected in the traditional environment. The experimental hypothesis is that Signal-to-noise Ratio (SNR) and the peak-latency convergence rates are statistically equivalent.

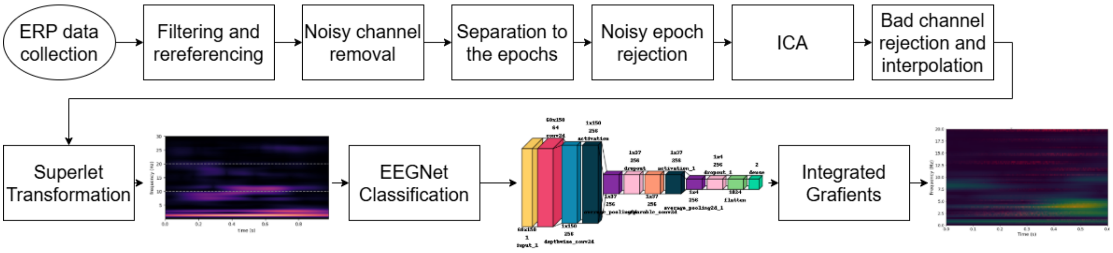

Figure 2: An illustration of the experimental pipeline, including data collection, pre-processing, epoching, transformation into superlets, CNN training, and XAI Integrated Gradients analysis.

To test this hypothesis, the Two One-Sided Tests (TOST) procedure was first selected to check the statistical equivalence of the signal-to-noise ratios of the four groups and the P3b peak latency convergence between ERPs collected in the VR and traditional lab environments. TOST performs two one-sided tests to check if the observed data is significantly larger than an equivalence boundary lower than zero (Δ_L) or substantially smaller than an equivalence boundary larger than zero (Δ_U) [21]. In addition, the Bland-Altman test [22] was used to verify the agreement and correlation between ERPs collected in the VR and the traditional lab environments and to find if there is a consistent bias. The SNR metrics used are: 1) log_{10} of the ratio of variance of signal to the variance of baseline, 2) log_{10} of the ratio of peak-to-peak amplitude of signal to the variance of baseline, 3) log_{10} of the ratio of root-mean-squared signal to root-mean-squared baseline and 4) log_{10} of the ratio of mean amplitude of the signal to the variance of baseline. SNR is measured in decibels, while peak latency is measured in milliseconds at each trial. The first derivative of each metric is averaged over trials to find its convergence rate. Even when ERP signals collected in different environments have statistically equal metrics, these environments may still influence the resulting waveform. Secondly, in addition to the TOST and the Bland-Altman test, an XAI technique is used to extract additional understanding of their differences. The two trained models are trained using the data collected in VR and traditional environments, and the Integrated Gradients method is applied. For the neural network $F : R^n \to [0, 1]$, where $x \in R^n$ is actual input and $x' \in R^n$ is some baseline input (black image or random noise), integrated gradients are found by cumulating the gradients along the straightline path from baseline to the input as the path integral. To find the integrated gradient along the i^{th} dimension for an input x and baseline x', equation 1 is used:

$$IntegratedGrads_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\delta F(x' + \alpha \times (x - x'))}{\delta x_i} \delta\alpha \tag{1}$$

Where $\frac{\delta F(x)}{\delta x_i}$ is the gradient of $F(x)$ along the i^{th} dimension [14].

4. Experimental results and discussion

The participant-wise averages of the signal recorded are plotted on the figure 3, first column. The green line is for the signals collected in the traditional lab environment, and the blue is for VR-based. The plot shows that the shape of the P3b signal is very clear and similar for the target events (top). However, the non-target events are visibly different after the peak area (after 0.3 seconds, bottom). This is due to the significantly lower nontarget P3b signal amplitude than the target one, making minor noise look significant. It is also important to mention that the averaging did not remove the consistent signal of the 10 Hz frequency. It is very obvious both at the averaged EEG signal and at most of the superlets: for example, fig. 3 VR-based env target (col. 3, top) and both target and non-target for the traditional environment, fig. 3 (col. 4). It means that it is some sort of signal going either from the natural EEG signal or from the equipment, and time-locked to the event in some way. The Signal-to-noise ratios and peak latencies for raw and preprocessed signals are statistically different (around 10% to 15% of groups have TOST p-values below 0.05). Still, they change very similarly, and the convergence rate for all groups is statistically equivalent - 83% - 94% of preprocessed datasets have TOST p-value below 0.05 and 91% - 100% of raw signals. Bland-Altman plots (an example in figure 3, col. 2) show a very high agreement with 90-95% of dots falling in $\pm 1.96SD$, for all the participants. The mean difference fluctuates around 0, slightly higher or lower for different groups. The AUC score used in [19] reached 0.72 for VR and 0.70 for the superlets associated with screen-based data. Classification results were separated into true-positives, true-negatives, false-positives and false-negatives, considering target ERPs as positives. Then, the Integrated Gradients method was run for each, and the results compared pairwise (VR vs screen) to facilitate the visual interpretation.

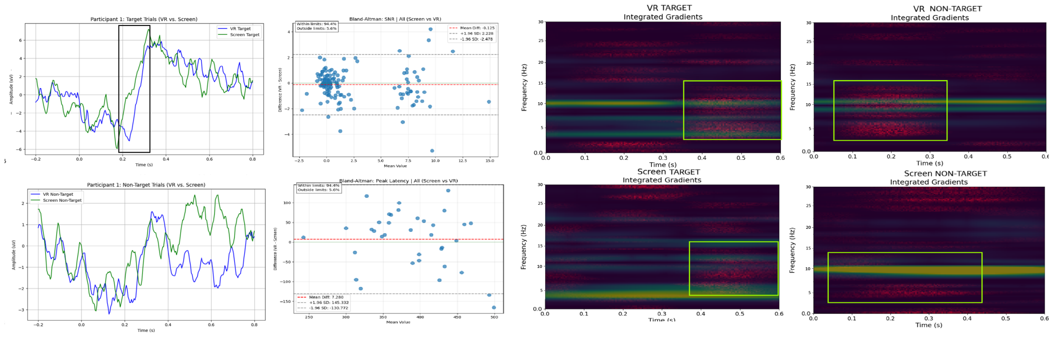

Figure 3: Averaged ERP-signals for 1 of the participants (col.1) and Bland-Altman plots for SNR and P3b peak latencies comparing convergence rates of the same metrics for the same participants (col. 2) for the ERP signals collected in the VR environment vs collected in the traditional lab environment, and Pairwise comparison of Integrated Gradients, target (col.3) and non-target (col.4)

A visual inspection of the figure 3 shows that the model trained to classify ERPs collected in VR and the model trained with natural lab environment signals have similarities, scalograms and integrated gradients. Such cases are consistent across participants. The interpretation of the integrated gradients reveals that both VR-specific and natural environment-specific models tend to focus on low (below 10 Hz) frequencies after the P3b peak for the target events and before for the non-target events. This follows the expectation that the P3b signal's peak for the target is large enough compared to the non-target events, where the peak amplitude is much lower (1 vs 6 μvolts on the fig. 3). However, the visualisation of IG shows that this effect is much stronger in the VR-associated signals (3^{rd} and 4^{th} columns on the fig. 3). Considering that the signals are the same, the noise is random, the signal-to-noise ratios are statistically equivalent, and preprocessing pipelines are the same for both environments, it is safe to assume that the aforementioned effect variation is created by the difference between the signals themselves. The longer attention is situated on the peak: 0.1-0.3 sec for VR vs 0-0.4 sec (last two images of the first row on the fig. 3) for traditional environment, which support the higher level of engagement in VR environments as it was reported by [23]. Also, as pointed out in [4], data collection in the VR environment caused higher artefact levels, the model successfully overcomes it by paying attention to a broader range of frequencies, as seen in the figure (1^{st} column on the fig. 3)

In conclusion, experimental findings show that the ERP signals collected in VR are slightly different in the noise distribution and the nature of the factors that decrease ERP measure quality compared to those from traditional environments. However, the signals provide the same amount of information about the ERP as is visually inspectable from the IGS. Future work includes, but is not limited to, extending the application of Integrated Gradients with GradCAM and SHAP to achieve further insights. Perturbation methods are envisioned as in [24], especially to understand the impact of the removal of the 10Hz noise on the classification accuracy of models. In line with this, alternative superlet parameters like the number of cycles are planned to localise ERP over time better.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] T. W. Picton, D. T. Stuss, The component structure of the human event-related potentials, Progress in brain research 54 (1980) 17–49.

[2] S. Kim, S. Lee, H. Kang, S. Kim, M. Ahn, P300 brain–computer interface-based drone control in virtual and augmented reality, Sensors 21 (2021) 5765.

[3] M. Simões, C. Amaral, P. Carvalho, M. Castelo-Branco, Specific eeg/erp responses to dynamic facial expressions in virtual reality environments, in: The International Conference on Health Informatics, Springer, 2014, pp. 331–334.

[4] V. Marochko, R. Reilly, R. McDonnell, L. Longo, A survey on the application of virtual reality in event-related potential research, in: A. Holzinger, P. Kieseberg, A. M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, Springer International Publishing, Cham, 2022, pp. 256–269.

[5] V. J. Harjunen, I. Ahmed, G. Jacucci, N. Ravaja, M. M. Spapé, Manipulating bodily presence affects cross-modal spatial attention: A virtual-reality-based erp study, Frontiers in human neuroscience 11 (2017) 79.

[6] C. D. Garduno Luna, Feasibility of Virtual and Augmented Reality Devices as Psychology Research Tools: A Pilot Study, Ph.D. thesis, UC Santa Barbara, 2020.

[7] J.-P. Tauscher, F. W. Schottky, S. Grogorick, P. M. Bittner, M. Mustafa, M. Magnor, Immersive eeg: Evaluating electroencephalography in virtual reality, in: 2019 Conference on Virtual Reality and 3D User Interfaces, IEEE, 2019, pp. 1794–1800.

[8] C. S. Herrmann, S. Rach, J. Vosskuhl, D. Strüber, Time–frequency analysis of event-related potentials: a brief tutorial, Brain topography 27 (2014) 438–450.

[9] D. Borra, E. Magosso, Deep learning-based eeg analysis: investigating p3 erp components, Journal of Integrative Neuroscience 20 (2021) 791–811.

[10] L. Dubreuil-Vall, G. Ruffini, J. A. Camprodon, Deep learning convolutional neural networks discriminate adult adhd from healthy individuals on the basis of event-related spectral eeg, Frontiers in neuroscience 14 (2020) 251.

[11] H. Moujahid, B. Cherradi, M. Al-Sarem, L. Bahatti, A. B. A. M. Y. Eljialy, A. Alsaeedi, F. Saeed, Combining cnn and grad-cam for covid-19 disease prediction and visual explanation., Intelligent Automation & Soft Computing 32 (2022).

[12] G. Vilone, L. Rizzo, L. Longo, A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence, in: The 28th Irish Conf. on Artificial Intelligence and Cognitive Science, volume 2771 of *CEUR Workshop Proceedings*, 2020, pp. 85–96.

[13] T. Ahmed, L. Longo, Examining the size of the latent space of convolutional variational autoencoders trained with spectral topographic maps of eeg frequency bands, IEEE Access 10 (2022) 107575–107586.

[14] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[15] Y. Kawai, K. Tachikawa, J. Park, M. Asada, Compensated integrated gradients for reliable explanation of electroencephalogram signal classification, Brain Sciences 12 (2022) 849.

[16] T. Donoghue, B. Voytek, Automated meta-analysis of the event-related potential (erp) literature, Scientific Reports 12 (2022) 1867.

[17] E. S. Kappenman, J. L. Farrens, W. Zhang, A. X. Stewart, S. J. Luck, Erp core: An open resource for human event-related potential research, NeuroImage 225 (2021) 117465.

[18] H. Nolan, R. Whelan, R. B. Reilly, Faster: fully automated statistical thresholding for eeg artifact rejection, Journal of neuroscience methods 192 (2010) 152–162.

[19] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, B. J. Lance, Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces, Journal of neural engineering 15 (2018) 056013.

[20] V. V. Moca, H. Bârzan, A. Nagy-Dăbâcan, R. C. Mureșan, Time-frequency super-resolution with superlets, Nature communications 12 (2021) 337.

[21] K. R. Murphy, B. Myors, Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model., Applied Psychology 84 (1999) 234.

[22] J. M. Bland, D. G. Altman, Measuring agreement in method comparison studies, Statistical methods in medical research 8 (1999) 135–160.

[23] T. Baumgartner, L. Valko, M. Esslen, L. Jäncke, Neural correlate of spatial presence in an arousing and noninteractive virtual reality: an eeg and psychophysiology study, CyberPsychology & Behavior 9 (2006) 30–45.

[24] M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: A survey, Pattern Recognition Letters 150 (2021) 228–234.