An application of Integrated Gradients for the analysis of the P3b event-related potential on a convolutional neural network trained with superlets

Vladimir Marochko[*,†], Luca Longo[†]

Artificial Intelligence and Cognitive load Research Lab, Technological University Dublin, Republic of Ireland.

Abstract

Event-related potentials (ERPs) are specific transient fluctuations in the brain's electrical field induced by the presentation of visual or auditory stimuli. They are time-locked and can help analyse neural activity in many applications. One of these, the P3b, elicited in the decision-making process, unlike other ERPs, is not linked to the physical characteristics of a stimulus but rather to a person's reaction to it. P3b ERPs can be evoked when completing a task with the oddball paradigm and have only slightly different waveforms when the target and non-target stimuli are presented. Identifying differences in these waveforms can help investigate dysfunctions in sensory and cognitive processing, among other applications. However, such identification is not straightforward because these differences are subtle and neural data contains many artefacts. Recently, deep learning has been used in addition to traditional methods to automatically learn the high-level features associated with target and non-target stimuli, discarding artefacts and supporting their discrimination. Unfortunately, even if powerful in creating discriminative models, they are regarded as black boxes because their inferential capacity is opaque and obscured. This research builds on this gap and explores an application of Integrated Gradients, a method developed within eXplainable Artificial Intelligence, to interpret a convolutional deep neural network, trained with superlets, specific time-frequency super-resolution of single-channel EEG signals, for discriminating target and non-target neural responses for an oddball task.

Keywords

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

1. Introduction

Event-related potentials (ERP) are specifically shaped fluctuations in brain electromagnetic activity that are time-locked to specific events, such as stimuli, responses, or decisions. They are produced in different brain areas and have various shapes depending on the nature of these events. They represent clusters of activation of neurons involved in the reaction to such events [1]. These potentials are strong enough to be recorded non-invasively with electroencephalography and can provide information about various cognitive processes [2], diagnostics of different neural [3], psychological conditions [4]. ERPs evoked by events of different natures are referred

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

[*] Corresponding author.

[†] These authors contributed equally.

vladimir.marochko@tudublin.ie (V. Marochko); luca.longo@tudublin.ie (L. Longo)
0000-0001-9766-6420 (V. Marochko); 0000-0002-2718-5426 (L. Longo)

to as components. In turn, components are defined by the perception source of an event or the mental process connected with it and of different natures [5]. In terms of the type of processing needed, ERP components have been found to appear in response to error detection and correction [6], pattern mismatching [7], and face expressions recognition [8]. Each component covers the different variety of ERP responses caused by its specific type of event inside its component [9]. The component chosen for investigation, one with great interest commonly researched in neuroscience, is the P300 component, often referred to as P3b. This is often evoked by the oddball task with two different groups of stimuli, in which a participant has to define whether the stimulus belongs to the smaller target group or the larger non-target one. The less frequent the target stimuli, the higher the difference between the peak amplitudes of the corresponding ERP, and the easier it is to classify signals [2]. Regression analysis [10], polynomial models trained with support vector machines [11] or boosting methods [12] were all adopted for classifying ERP signals. More recently, deep learning [13] and specifically convolutional neural networks were considered [14, 15]. However, the internal structure of the trained models with these techniques is opaque and hidden from human perception [16]. This research explores how the Integrated Gradient (IG) attribution method, developed within explainable Artificial Intelligence (XAI) [17], can improve the interpretability of a convolutional neural network and improve human trust towards AI-based solutions [18]. This network is trained with single-channel superlets to classify signals associated with target and non-target and facilitate the extraction of the ERP component. These are specific time-frequency super-resolution representations, similar to scaleograms, that can be visually inspected [19]. The remainder of the manuscript includes a presentation of relevant work in section 2, followed by a description in section 3 of the design of an experiment and its methods. Eventually, section 4 presents a critical discussion on the impact of IG for analysing the trained neural network in the context of P3b detection.

2. Related work

The P3b event-related potential is one of the most studied among the others ERPs [2] P3b scalp distribution is defined as the amplitude change over the midline electrodes (Fz, Cz, Pz placed according to the International 10/20 System), which typically increases in magnitude from the frontal to parietal electrode sites [20]. This component has a sharp positive voltage spike that occurs 300 ms after the stimulus is received, which causes such naming. P3b appears to be evoked by the oddball stimuli detection tasks [21, 22]. Many methods are traditionally used within neuroscience to improve the extraction of the P3b component. These include frequency decomposition [23], discriminant score [10], independent component analysis [24]. Recently, leveraging the success of machine learning in various application areas, deep learning methods have been used in computational neuroscience for multiple tasks employing electroencephalographic data, including the detection of schizophrenia [25], or Parkinson's in resting-state EEG [26]. One problem with extracting P3b components is the similarity in neural responses to target and non-target stimuli. For this reason, deep learning was used to improve such extraction [14] because of its capacity to learn high-level features from complex EEG data. However, even if it usually leads to developing robust and accurate classifiers in many applications, this is not the case in ERP research [27]. To improve P3b detection, EEG segments are usually transformed in the time-frequency domain, either leading to spectrograms [22] or scaleograms [28] Subsequently, such representations are used with Convolutional Neural Networks to train discriminative models [29, 28]. However, although these models' performance is reasonable, they are regarded as black boxes because it is difficult to interpret their inferential mechanisms and understand the high-level features they have learned to discriminate the dependent classes. In other words, the internal structure of these models is hidden from human perception [16]. The field of explainable Artificial Intelligence (XAI) actively explores different methods for interpreting black boxes [17]. Applying XAI methods to data-driven trained models for discriminating target and non-target classes, as in the case of the P3b, can help understand what high-level features contribute exactly to discriminating these classes, facilitating the understanding of the nature underlying their differences [30]. One of the XAI methods that is simple and suitable for such tasks is the Integrated Gradients [31]. It is used to visualise the influence of different input parts on a neural network's class mapping [32]. In detail, the method is based on computing the model's gradients on the path from the baseline input to the actual input and then accumulating these gradients. This axiomatic attribution method is simple to implement and replicate because it does not require almost any modification of the underlying network it aims to interpret [31]. Recently, this has been applied with EEG data for ERP research [33]. This trend inspired this study, and a novel empirical experiment, as described in the next section, was designed to merge deep learning, trained with superlets, unique time-frequency representations of EEG data, and Integrated Gradients from XAI.

3. Experiment design and methods

An empirical study was designed to support the identification of the P3b ERP component by employing deep learning with superlets and Integrated gradients, as depicted in Fig. 1).

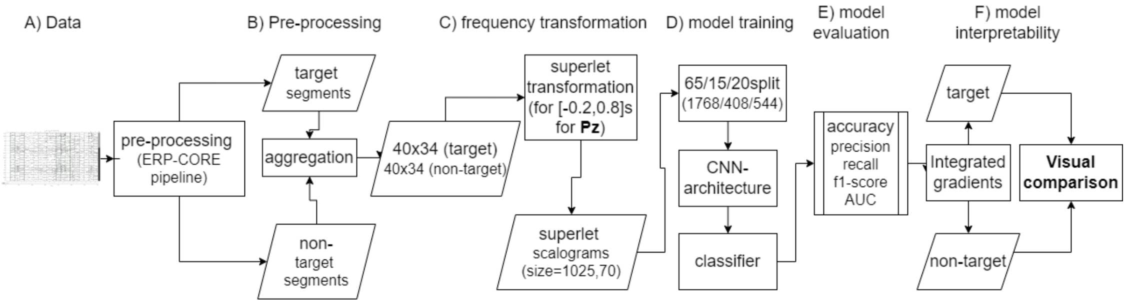

Figure 1: The design of the research experiment

Phase A: Data selection - The ERP-core project was selected, a freely available set of resources on different ERP paradigms, experiments, electroencephalographic data and functions, such as pre-processing steps, to support methods applicability and results reproducibility [34]. The P3b (Active Visual Oddball Paradigm) was selected, containing data from 40 participants. The experiment consists of 5 trials, with a short break between them. For each trial, one of the 5 letters (A, B, C, D, E) is randomly chosen as a target letter; the rest are non-targets (standards).

Then, a single letter is shown 40 times for each trial, chosen randomly from a pool. Eight letters are the chosen targets, and the remaining 32 are randomly selected from the non-target. The letter is shown for 0.2 seconds. Subsequently, a black screen appears for 1200-1400 ms, with a white dot in the centre. When the letter is shown, a participant must recognise whether it is the target and press the up arrow on the keyboard or the down arrow otherwise. EEG data was recorded during the entire experiment along with a timestamp for each shown letter, and what letter was the target and which one was shown [34].

Phase B: pre-processing - Makoto's preprocessing pipeline is a widely used protocol in ERP studies to clear the noise, remove artefacts, and extract the signal. Inspired by this, EEG signals were preprocessed by shifting the epochs to 28 ms to compensate for the screen latency, applying a band-pass filter (0.01-45 Hz) and performing Independent Component Analysis (ICA) to remove artefacts. Only data from the Pz channel is chosen because the P3b signal at this location is the strongest [34]. The ERP-Core team excluded data from 6 out of 40 participants because it was considered too noisy due to too strong artefacts. The EEG data of the remaining 34 participants was preserved, split into intervals around stimuli presentation, precisely 200 milliseconds before and 800 milliseconds after it.

Phase C: Time-frequency super-resolution - State-of-the-art deep learning approaches have employed spectrograms or scaleograms for time-frequency resolution of single-channel EEG signals as training instances [29, 28]. However, due to the Heisenberg-Gabor uncertainty principle, finite oscillation transients are difficult to localise simultaneously in time and frequency. Classical estimators, including the short-time Fourier or the continuous-wavelet transforms, optimise temporal or frequency resolution or find a suboptimal tradeoff. Plausibly and consequently, this is why deep learning-based models trained with single-channel EEG time-frequency representations exhibited a sub-optimal performance [14]. To counteract this limitation, this study employed superlets [19]. A superlet (SL) is a finite set of o (order) wavelets spanning a range of c different multiple bandwidths (cycles) at the same central frequency f. These are combined geometrically to maintain the good temporal resolution of single wavelets and gain frequency resolution in upper bands. The parameters for the superlets transformation were chosen: sampling frequency of 1024 Hz, band-pass filter within 0.01 to 30 Hz, the base number of cycles parameter is 5, and the order spanned across the frequencies of interest is 8. As in the case of scaleograms, a superlet can be shaped and visualised as a matrix.

Phase D: Model training - All the intervals around each stimulus (-200ms, +800ms) at the Pz channel, with 1024 hertz, are converted to superlets. These superlets are normalised from 0 to 1 and labelled with [1, 0] for the target stimuli and [0, 1] for non-target stimuli. Subsequently, these instances are separated into training, validation, and test sets with a 65/15/20% ratio (1768, 408, and 544 instances). Due to the unbalanced nature of the data, since non-target instances are significantly higher than target instances, undersampling was performed randomly to the majority class (non-target) to balance the training material evenly for each participant. The classifier devised is a deep convolutional neural network with five hidden convolutional layers, a Relu activation function, a 0.1 dropout rate, 1024 filters in the first four layers, and 512 in the fifth. The kernel sizes are 3, 5, 7, 11 and 13. The flattening layer produces vectors. One dense hidden layer has 512 hidden units, Relu activation and the Dense output layer has two output units with the softmax activation function. The model is trained with the Adam Optimiser. The loss function used is categorical cross-entropy. Early stopping with a patience of 10 and a minimal delta of 1e-4 is applied to avoid overfitting.

Phase E: Model evaluation - The model is evaluated using accuracy, precision, recall, the F1 score, and the ROC Area Under the curve (AUC).

Phase F: Model Interpretability - This research uses integrated gradients to explain how the trained model chooses the class during classification. The one-hot version of the classification method is chosen to have separate firing neurons for different classes. For the target (TRUE) and non-target classes (FALSE), integrated gradients can be computed and demonstrated separately for each class. Finding integrated gradients is based on creating a random baseline for the image and then computing model gradients on the path concerning the actual data superlets. These were grouped by true positives (TP: actual targets), true negatives (TN: actual non-targets), false positives (FP: non-targets recognised as targets), and false negatives (FN: targets recognised as non-targets) to facilitate interpretation and understanding of why the model learnt the mapping.

4. Results and Discussion

The trained model exhibited 0.6342 accuracy, 0.6346 precision, 0.6324 recall and ROC-AUC 0.659. In detail, the TP rate is 0.634, the TN is 0.633, the FP is 0.365, and the FN is 0.366 While these metrics might appear low, building an automatic binary classifier for discriminating targets from non-target brain responses is extremely difficult, given their subtleties. Anyway, such results are higher when compared to a similar work [35], thus demonstrating an improvement. A visual analysis of the averaged time series, as depicted in figure 2, confirms that the event-related potentials P3b were higher for true-positive (TP, blue) than the other cases.

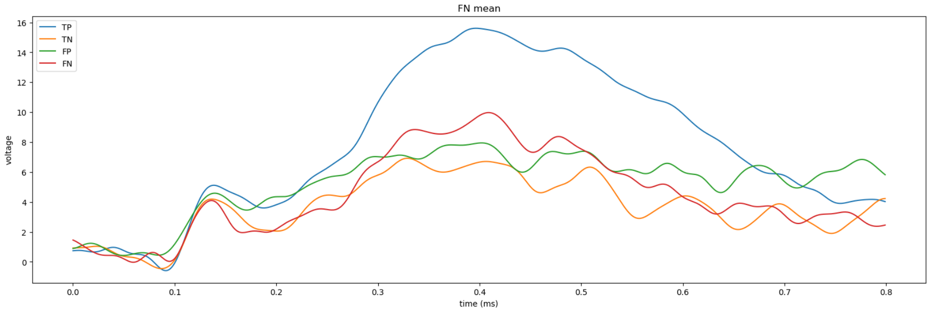

Figure 2: Averaged signals in different groups - the baseline noise is very high in misclassified ones).

Given this, it is intuitive to expect that the integrated gradients associated with the TPs exhibit some characteristics different from those associated with the other cases (FP, FN, TN). Figure 3 marks in red the areas where the integrated gradients of the trained model are the highest overlayed over the average of the superlets from the respective group to mark the most important parts. Noticeably, on the one hand, the true positives (TPs) have a clear, well-defined region of integrated gradients between 0.15 and 0.45 ms in the delta band (area 1), exactly where neural activity associated with the P3b erp is expected. As expected, it also highlights a time transition of activity from this region of integrated gradients to another (read 2) in the theta band. On the other hand, the average superlet associated with false positives (FP) deviates from this configuration by showing an additional region of activity in the beta band (area 3). This new area contains integrated gradients that are not present for TP, but they are weaker than those in areas 2 an 3, suggesting this might be EEG noise. Removing this area, probably the 0.365% of false positives can turn into true positives. Regarding true negatives (TNs), evident activity can be seen in the low part of the delta band across the timeline (area 4), suggesting that, successfully, no actual neural peak of activity can be spotted post-stimulus. Unfortunately, the activity of the false negatives (FNs) is similar to this, and no significant differences in integrated gradients can be spotted, hence their incorrect classification. However, examining Figure 2, the average of P3b ERPs for FN (in red) is higher than that of the TN (orange), suggesting why the model made a wrong classification. Again, perhaps by better pre-processing the raw EEG signals, noise can be mitigated, and a higher accuracy can be achieved. Overall, this research study demonstrated a possible application of integrated gradients on the superlets associated with brain responses post-stimulus to help spot the characteristics of P3b event-related potential and support visual analysis.

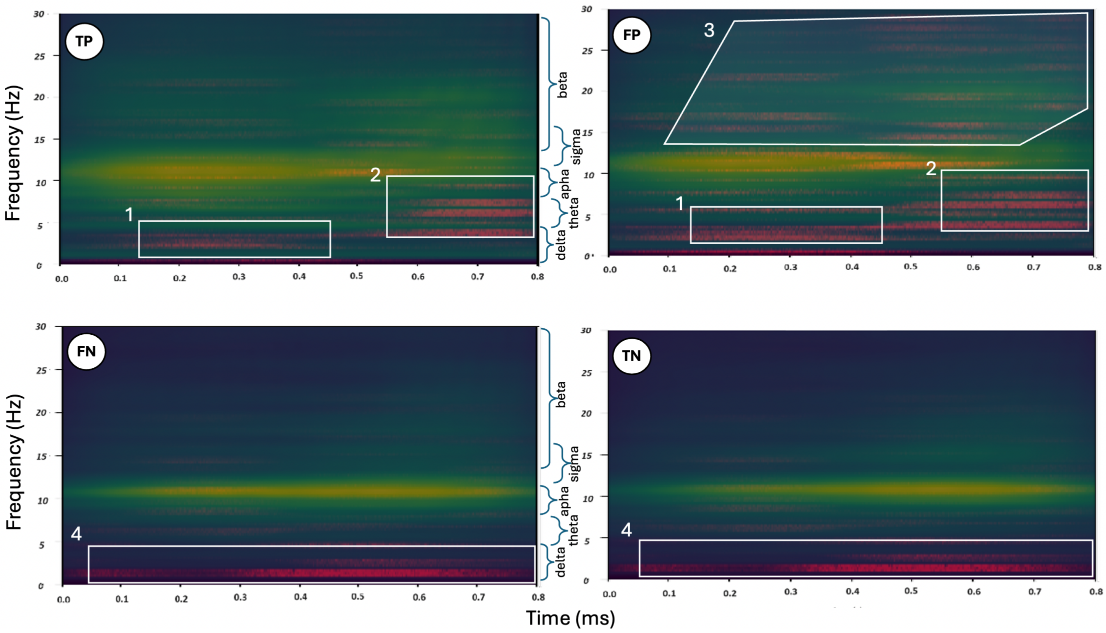

Figure 3: Integrated gradients associated with the true-positive, false-positive, false-negative, and true-negative on the training data (target class=positive, non-target=negative) overlapped on the averages of the superlets in their respective groups.

Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission

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. J. Luck, Event-related potentials., APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics. (2012). doi:10.1037/13619-028.

[3] M.-H. Ahn, J. H. Park, H. Jeon, H.-J. Lee, H.-J. Kim, S. K. Hong, Temporal dynamics of visually induced motion perception and neural evidence of alterations in the motion perception process in an immersive virtual reality environment, Frontiers in neuroscience 14 (2020).

[4] G. Hajcak, J. Klawohn, A. Meyer, The utility of event-related potentials in clinical psychology, Annual Review of Clinical Psychology 15 (2019) 71–95.

[5] Y. Yokota, Y. Naruse, Temporal fluctuation of mood in gaming task modulates feedback negativity: Eeg study with virtual reality, Frontiers in human neuroscience 15 (2021) 246.

[6] L. Koban, G. Pourtois, R. Vocat, P. Vuilleumier, When your errors make me lose or win: event-related potentials to observed errors of cooperators and competitors, Social Neuroscience 5 (2010) 360–374.

[7] P. Paavilainen, The mismatch-negativity (mmn) component of the auditory event-related potential to violations of abstract regularities: a review, International journal of psychophysiology 88 (2013) 109–123.

[8] L. Kirasirova, A. Zakharov, M. Morozova, A. Y. Kaplan, V. Pyatin, Erp correlates of emotional face processing in virtual reality, Opera Medica et Physiologica 8 (2021) 12–19.

[9] S. Sur, V. K. Sinha, Event-related potential: An overview, Industrial psychiatry journal 18 (2009) 70.

[10] K. C. Squires, C. Wickens, N. K. Squires, E. Donchin, The effect of stimulus sequence on the waveform of the cortical event-related potential, Science 193 (1976) 1142–1146.

[11] U. Hoffmann, J.-M. Vesin, T. Ebrahimi, Spatial filters for the classification of event-related potentials, Proceedings of ESANN 2006 (2006).

[12] U. Hoffmann, G. Garcia, J.-M. Vesin, K. Diserens, T. Ebrahimi, A boosting approach to p300 detection with application to brain-computer interfaces, in: 2nd International IEEE EMBS Conference on Neural Engineering, 2005., IEEE, 2005, pp. 97–100.

[13] W. Buntine, Machine learning after the deep learning revolution, Frontiers of Computer Science 14 (2020) 1–3.

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

[15] S. A. S. Bellary, J. M. Conrad, Classification of error related potentials using convolutional neural networks, in: 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2019, pp. 245–249.

[16] W. J. von Eschenbach, Transparency and the black box problem: Why we do not trust ai, Philosophy & Technology 34 (2021) 1607–1622.

[17] L. Longo, et. Al., Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301.

[18] P. Dondio, L. Longo, Trust-based techniques for collective intelligence in social search systems, in: Next generation data technologies for collective computational intelligence,

Springer, 2011, pp. 113–135.

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

[20] J. Polich, Updating p300: an integrative theory of p3a and p3b, Clinical neurophysiology 118 (2007) 2128–2148.

[21] T. W. Picton, et al., The p300 wave of the human event-related potential, Journal of clinical neurophysiology 9 (1992) 456–456.

[22] A. Mussabayeva, Z. Ermaganbet, P. K. Jamwal, M. T. Akhtar, Event-related spectrogram representation of eeg for cnn-based p300 speller, in: 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE, 2021, pp. 410–415.

[23] R. Naafanen, The role of attention in auditory by event-related potentials and other brain measures of cognitive function, Behavioral and Brain Sciences 13 (1990) 201–288.

[24] T. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, T. Sejnowski, Independent component analysis of single-trial event-related potentials, in: Proc. ICA, volume 99, 1999, pp. 173–179.

[25] N. Grover, A. Chharia, R. Upadhyay, L. Longo, Schizo-net: A novel schizophrenia diagnosis framework using late fusion multimodal deep learning on electroencephalogram-based brain connectivity indices, IEEE Transactions on Neural Systems and Rehabilitation Engineering 31 (2023) 464–473. doi:10.1109/TNSRE.2023.3237375.

[26] U. Lal, A. V. Chikkankod, L. Longo, Fractal dimensions and machine learning for detection of parkinson’s disease in resting-state electroencephalography, Neural Computing and Applications (2024) 1–24.

[27] A. R. Marathe, A. J. Ries, V. J. Lawhern, B. J. Lance, J. Touryan, H. Cecotti, The effect of target and non-target similarity on neural classification performance: a boost from confidence, Frontiers in neuroscience 8 (2015) 135029.

[28] Q. Xin, S. Hu, S. Liu, L. Zhao, Y.-D. Zhang, An attention-based wavelet convolution neural network for epilepsy eeg classification, IEEE Transactions on Neural Systems and Rehabilitation Engineering 30 (2022) 957–966.

[29] B. Mandhouj, M. A. Cherni, M. Sayadi, An automated classification of eeg signals based on spectrogram and cnn for epilepsy diagnosis, Analog integrated circuits and signal processing 108 (2021) 101–110.

[30] W. Guo, Explainable artificial intelligence for 6g: Improving trust between human and machine, IEEE Communications Magazine 58 (2020) 39–45.

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

[32] W. Ma, Y. Li, X. Jia, W. Xu, Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4630–4639.

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

[34] 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.

[35] J. P. A. Nogueira, Exploring transfer learning techniques for P300-based Brain Computer Interfaces, Master’s thesis, University of Coimbra, 2022.