Detecting Internal and External Attention in Virtual Reality: A Comparative Analysis of EEG Classification Methods

Francesco Chiossi
LMU Munich
Munich, Germany
francesco.chiossi@um.ifi.lmu.de

Changkun Ou
LMU Munich
Munich, Germany
research@changkun.de

Felix Putze
University of Bremen
Germany
fputze@uni-bremen.de

Sven Mayer
LMU Munich
Munich, Germany
info@sven-mayer.com

Abstract

Future VR environments envision adaptive and personalized interactions. To this aim, attention detection in VR settings would allow for diverse applications and improved usability. However, attention-aware VR systems based on EEG data suffer from long training periods, hindering generalizability and widespread adoption. This work addresses the challenge of person-independent, training-free VR BCI classifying internal and external attention in VR. We compared the performance of four classifiers on an EEG dataset (N=24) featuring internal and external attention labeled classes. With the goal of online adaptation, we tested overall accuracy, different window lengths of the data, and training split to optimize the trade-off between window length and classification accuracy. Our results show that models using a complete EEG band combination consistently achieve the highest accuracy, with Linear Discriminant Analysis particularly benefiting from full-band data. The window length impacts most models' performance with short windows. LDA achieved optimal accuracy around 6.3 seconds, SVM and NN around 6.5 and 6 seconds, respectively, and RF reached stability at 6 seconds. Lastly, increasing training data ratios improved accuracy gains consistently across models. We discuss the potential of machine learning to model EEG correlates of internal and external attention as online inputs for adaptive VR systems.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

EEG, Virtual Reality, Adaptive Systems, Attention

ACM Reference Format:
Francesco Chiossi, Changkun Ou, Felix Putze, and Sven Mayer. 2024. Detecting Internal and External Attention in Virtual Reality: A Comparative Analysis of EEG Classification Methods . In *International Conference on Mobile and Ubiquitous Multimedia (MUM '24), December 1–4, 2024, Stockholm, Sweden.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3701571.3701579

MUM '24, December 1–4, 2024, Stockholm, Sweden

ACM ISBN 979-8-4007-1283-8/24/12
https://doi.org/10.1145/3701571.3701579

1 Introduction

Virtual Reality (VR) technology has witnessed significant advancements in recent years, with its applications through various domains, including gaming [19], healthcare [45], and training [20]. As VR evolves, the focus has shifted towards developing adaptive systems intelligently adapting to users' states in real-time [3]. This paradigm shift is driven by the recognition that personalized and dynamically tailored experiences are crucial to enabling more diverse VR interactions [14]. Adaptive interaction has become central to designing engaging VR experiences.

To enable adaptive VR systems, the emerging physiological computing perspective provides a promising approach [23]. Physiological computing leverages the understanding that human physiological signals can serve as a valuable source of information for interaction and adaptation. Using physiological data as input for interaction, we can gain insights into users' cognitive and affective states, tailoring the VR experience to their current goals.

When engaged in VR interactions, we might be exposed to a wide range of stimuli, varying in intensity and frequency, requiring externally-directed attention and, at the same time, be engaged in productive or cognitive tasks that might require memory recall [56], and mental arithmetic [1], i.e., internally-directed attention. The relevance of distinguishing between internal and external attention became evident across different VR tasks [42, 50, 53]. Consequently, levels of attention can fluctuate, influenced by intrinsic and extrinsic factors and events in the VR environment that could disrupt such processes. Whether internal or external attention states can be preferable depending on the VR interaction, attentional mechanisms play a central role in directing focus toward relevant information while suppressing what is perceived as irrelevant. The challenge lies in effectively suppressing sensory input, particularly when it becomes more perceptually salient [24]. The controlled nature of VR enables the adaptation of digital content, offering the flexibility to manipulate and control the appearance of potentially distracting or task-irrelevant information. This adaptability allows for the display of virtual content that can effectively respond to the user's attentional state and align with the objectives of the application. Conversely, if we display relevant information in the VR environment, but users direct their attention internally, i.e., in a mind-wandering state, adaptation might increase the saliency of external stimuli to aid users in maintaining their concentration on the task.

In this work, we evaluate machine learning-based approaches for the implicit classification of internal and external attention states based on electroencephalographic (EEG) features derived from the dataset of Chiossi et al. [11]. Verbalizing or recalling the attentional

state directly from participants can be challenging, as it may involve subconscious or forgotten aspects. Thus, we investigated the feasibility of internal and external attention implicit detection to address this limitation.

Secondly, we target the overall problem of model generalizability for neuroadaptive technology [35]. In neuroadaptive VR applications, achieving high classification accuracy often involves training the classifier on data specific to each individual, resulting in personalized models that are hard to generalize across participants. The training process typically requires explicit labeling of recorded data in time-consuming sessions. However, if the data could be person-independent, it would remove the need for extensive data collection and enable the system to be used without prolonged calibration. By making the system person-independent, it would be possible to pre-train the classifier using data from multiple users. This approach would increase the size of the training dataset, reduce bias, and mitigate the risk of overfitting on limited training samples. Ultimately, person independence offers the potential for more efficient and widely applicable neuroadaptive VR systems.

With those goals in mind, we systematically evaluated different model accuracy over different EEG features, window segmentation, and applicability of person-independent classification for future online use. Based on previous work in Augmented Reality (AR) [62], we compared a Linear Discriminant Analysis (LDA) model, and Support Vector Machine (SVM) to non-linear algorithms, i.e., Random Forest Classifier and a simple Neural Net.

We make the following contributions: we conducted a systematic analysis to identify the optimal feature set (I) for classifying internal and external attention states in VR environments. Second, we investigated the influence of time window segmentation (II) on classification performance. Furthermore, we identified optimal time windows for attention detection based on each model (III). Four, we showed the different impacts of model parameters on accuracy and provided guidelines for their selection (IV). Lastly, we make our analysis approach and preprocessed datasets openly available (V) .

2 Related Work

Here, we review the relevance of internal and external attention in VR settings and summarize previous work in internal and external attention detection based on EEG features for adaptive interaction.

2.1 Internal and External Attention in Virtual Reality

When immersed in VR settings, users need to allocate their attention to both external stimuli and internal mental processes, as described in Chun et al.'s taxonomy [15], which differentiates between external attention directed towards the VR environment and internal attention focused on cognitive tasks and mental representations.

External attention encompasses allocating attention towards external stimuli, which can occur through top-down or bottom-up processes. Task demands and goal-oriented features drive top-down attentional control, whereas bottom-up attentional capture occurs when attention is involuntarily drawn to objects or events [26, 54]. On the other hand, internal attention involves processing and updating internal representations of information, including working and prospective memory or mental calculation [16, 49]. Internal attention is often guided by top-down processes such as goals or motivation.

External and internal attention have been investigated in VR research to study immersion and engagement [10, 55]. For instance, Magosso et al. [42] compared the attentional competition between external and internal allocation when engaged in a mental arithmetic task (internal attention) in an immersive VR environment (external attention). A realistic VR environment requires more external attention resources like those required when reading. This finding aligned with Ricci et al. [50], showing that exposure to a VR environment increased external attention compared to a relaxation state that recruits internal attention. Similarly, Dey et al. [18] found increased EEG alpha oscillations in parietal sites associated with external attention and a higher sense of presence in VR.

Attentional states also play a significant role in determining users' engagement with a specific task. For example, Katahira et al. [30] studied different flow experiences during an internal attention task (a mental arithmetic task) reporting that EEG correlates of external and internal attention could distinguish between states of overload, boredom, and flow. Lim et al. [37] compared internal attention and immersion in desktop and VR settings, where beta and alpha frequencies could discriminate between the two states.

Considering the relationship between internal and external attentional states and the VR experience, attention detection in VR would allow for enlarging VR scenarios' interaction space and user experience. By leveraging the insights gained from studying external and internal attention, researchers can develop adaptive systems to detect users' attentional states in VR. Next, we review existing approaches for attention detection in VR.

2.2 Attention Detection in Augmented and Virtual Reality

Attention detection in VR environments is crucial for improving user engagement and optimizing content, visualizations or interaction [14]. Thus, several studies explored the feasibility of internal and external attention detection in AR and VR settings, focusing on different adaptive strategies and feature sets based on eye-tracking and EEG.

In AR, Vortmann et al. demonstrated the effectiveness of simple machine learning techniques, particularly LDA, in distinguishing between internal and external attention states, pointing toward real-time attention assessment in AR applications [60]. They extended this work by integrating the LDA model into an attention-aware language translation application for AR mobile devices, confirming the practicality of attention detection in real-world settings [63]. Furthermore, Vortmann and Putze developed an attention-aware BCI using Steady-State Visually Evoked Potential (SSVEP) to reduce distractions, enhancing system usability through a nearest-neighbour classification [61]. These studies collectively highlight the potential for utilizing attention detection techniques to enhance user experience and mitigate distractions in AR environments employing machine learning approaches.

Instead, in VR settings, the most common approach employed rule-based adaptation. Ewing et al. [22] adjusted game difficulty by the higher-level goal of sustaining attention and motivation to the game-adapted sensory and challenge immersion in a VR

game. Their system architecture was based on variation threshold from baseline on EEG features over a 4s time window. Similarly, but with a focus on optimizing internal attention states, Chiossi et al. [11] adapted the visual complexity of visual distractors while participants were engaged in a Working Memory task, ultimately resulting in improved behavioral performance.

Souza and Naves [55] reviewed recent literature for VR attention detection via EEG features. First, they emphasized the importance of personalized machine learning models for attention detection. Second, they highlighted the need for high-resolution EEG techniques to capture brain connections and information about immersion, attention, and cognitive load. Various approaches were employed to study attention allocation under different immersion levels and cognitive loads; however, few examples of online applications of machine learning algorithms were investigated for online adaptation in VR environments. Thus, optimizing user interfaces based on attentional states has been primarily explored in AR. However, real-time assessment and adaptation based on users' attentional state still need to be improved in VR paradigms.

2.3 Research Gap

This study focuses on detecting internally and externally directed attention in VR. Previous research has explored the EEG mechanisms underlying internal and external attention in AR and VR. Still, there is a lack of studies investigating person-independent Brain-Computer Interfaces (BCIs) for distinguishing between these states in VR. Thus, the primary objective of this study is to establish an optimal and reliable approach for real-time classification without the need for prior classifier training in VR, based on previous work in AR for attention detection [62]. To achieve this, we formulate different research questions.

Firstly, we focus on the labeled EEG dataset collected by Chiossi et al. [11]. Following the recommendations by Lotte et al. [40], we integrate their LDA results with the classification performance of RF and SVM classifiers with a recursive feature selection approach. Additionally, we compare their performances against a Neural Network (NN). Prior studies have suggested NNs are more effective for EEG data in workload classification than linear algorithms [2, 17, 34]. For better comparability, a vanilla neural network is initially trained using the same feature set to determine if it exhibits any advantages with identical features. This approach would allow us to verify: **RQ1** Which are the best feature sets to classify external and internal attention in VR?

Next, we aim to investigate: **RQ2** Which is the optimal time window length for achieving high classification accuracy? Given the rapid fluctuations in attention, longer time windows may encompass periods of both internal and external states, potentially compromising accuracy [15]. Conversely, shorter epoch lengths might better estimate the current state but could reduce accuracy due to fewer data samples. This investigation also examines whether a higher number of time windows (and consequently, feature sets) used for training the classifier improves the accuracy of state prediction or if the overall length of the training data in seconds is more critical (**RQ3**). This inquiry will determine if segmenting short training periods into multiple windows is as effective as using fewer, longer windows over extended sessions.

Thus, addressing **RQ1** and **RQ2** leads us to reformulate and introduce a new **RQ3**: How does the size of training data influence the trade-offs between model accuracy and computational efficiency?

3 Dataset Processing

In this work, we analyze the EEG dataset from Chiossi et al. [11] labeled with their attentional state (internal and external). Their dataset encompassed EEG frequencies (delta, theta, alpha, beta, and gamma) averaged every 20 s in a Visual Monitoring task, recruiting external attention [59], and a visual Working Memory (WM) N-Back task, recruiting internal attention resources [12, 13, 33].

3.1 Experimental Tasks

In the original study, participants engaged with two types of tasks during the study: Visual Monitoring task and N-Back task.

3.1.1 External Attention - Visual Monitoring. In the Visual Monitoring task, participants were presented with a continuous stream of non-player characters (NPCs) at a rate of 334 NPCs per minute. They were instructed to visually track and follow NPCs that appeared in randomly assigned colors (blue, green, black, and red). This task was designed to engage external attention resources, requiring substantial visual processing.

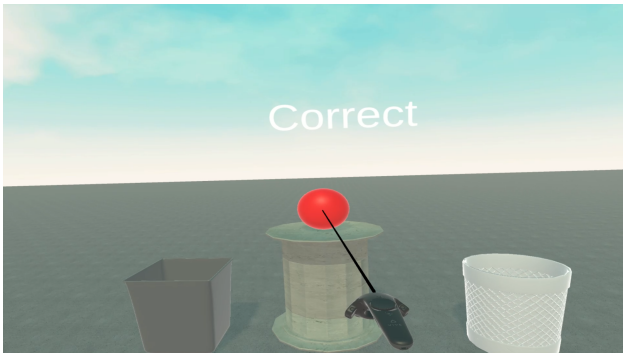
3.1.2 Internal Attention - N-Back Task. Participants performed the N-Back task in the Internal attention block (N=2). Participants were presented with a sequence of spheres positioned on a marble-like pillar. They had to place each sphere into one of two buckets on the left and right sides, respectively. The spheres could appear in four different colors (green, red, blue, and black) in a randomized sequence. The placement of each sphere depended on its color, and the sphere's color presented two steps before it. If the colors were the same, the participant had to place the sphere in the right bucket. If the colors differed, the participant had to put the sphere in the left bucket. New spheres would appear either after the current sphere was placed in one of the buckets or after a delay of 4 seconds.

3.2 EEG Features Extraction

We utilized MNE Python [25] for EEG analysis. The preprocessing pipeline involved a notch filter at 50 Hz, followed by a band-pass filter ranging from 1 to 70 Hz. The signal was then re-referenced to the common average reference (CAR) and subjected to Independent Component Analysis (ICA) using the Infomax algorithm. We employed the "ICLabel" MNE plugin [48] for automated classification and correction of ICA components.

The epoch features were derived from the Power Spectral Densities (PSDs) of different electrode groups based on prior work [11, 27, 38] via the Welch method. We extracted average powers for Delta (.5 - 4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), Beta (14-30 Hz), and lower Gamma-band (30-45 Hz). This resulted in following ROIs: Fz, F3, F4, F7, F8, Cz, C3, C4, Pz, P3, P4, Oz, O1, O2 for delta; Fz, F1, F2, F3, F4, FC1, FC2 for theta and beta; Pz, P1, P2, P3, P4, POz, PO3, PO4, Oz, O1, O2 for alpha; Fz, F1, F2, F3, F4, FC1, FC2 for beta; Fz, F3, F4, FT7, FT8, Cz, C3, C4, Pz, P3, P4, PO7, PO8, Oz for gamma. We employed extracted EEG frequencies over 20 different window lengths, varying from 1 second to 20 sec in steps of 1 sec.

(a) External Attention - Visual Monitoring



(b) Internal Attention - N-Back

Figure 1: Experimental Tasks. In the Visual Monitoring task (a), participants are exposed to a Stream of NPCs and tasked with tracking NPCs of a specified color by following them with their gaze. In the N-Back (b), participants interact with a sequence of spheres displayed on a pillar. They must sort each sphere into either the left or right bucket based on the color match to the sphere presented two steps prior (N=2). Spheres are placed in the left bucket if their color differs from the reference, and in the right bucket if the color matches.

4 Classification and Hyper-Parameter Optimization

In this study, we aimed to investigate the impact of different parameters, i.e., feature set, model, and training ratio, on classification accuracy. We selected LDA based on previous research as it has demonstrated effectiveness for binary classification tasks in VR and AR settings [11, 60, 64], together with an SVM [55]. Additionally, we included an RF classifier [40]. However, Appriou et al. [2] suggested that a neural network approach may yield better results for EEG classification. Thus, we also implemented a vanilla neural network.

We evaluate the predictions using accuracy using equally frequent internal and external classes and stratified test and training splits. The baseline chance level for correct class prediction was set at 50%. For hyper-parameter optimization, we performed a grid search with 8-fold cross-validation for all four classifier parameters: LDA, SVM, RF, and NN. We looked for the best internal model parameter configuration in the grid search.

We systematically varied the features, window length, and train-test split to address our research questions about how the data should be prepared for training. In the process, we shuffled the dataset and did a 5-fold bootstrapping for each training configuration to ensure hyperparameter robustness. For features, we used all 31 unique EEG band combinations of Delta, Theta, Alpha, Beta, and Gamma. We tested lengths of 1 to 20 seconds for window length in steps of 1 sec. For the train-test split, we tested 50%, 70%, and 90% of the participants in the train split; we used the remaining participants for testing. This resulted in $31 \times 20 \times 3 \times 5 = 9300$ runs for internal model hyperparameter turning, aka the above-described grid search.

This approach ensures that the model is both tailored to the specific characteristics of the dataset and robust against overfitting and underfitting, leading to reliable predictive performance when applied to the test data.

4.1 Linear Discriminant Analysis

We implemented the LDA model using the Python Scikit-learn toolbox [46]. We performed a grid search on the parameters `solver` and `shrinkage`, and `n_components`. For `solver`, we systematically tested *svd*, *lsqr*, and *eigen*. For `shrinkage`, we varied between *None*, and *auto* using the Ledoit-Wolf lemma. Shrinkage regularization is crucial for improving the robustness and accuracy of the LDA, particularly when dealing with datasets where the number of features may be comparable to, or exceed, the number of samples. Shrinkage helps stabilize the computation of the covariance matrix by adjusting it towards a scaled identity matrix, thus mitigating issues related to its singularity or near-singularity. For `n_components`, we tested 2 to 5 components and None.

4.2 Support Vector Machine

We employed the Support Vector Machine (SVM) classifier via the Python Scikit-learn toolbox [46]. We performed a grid search on the parameters `C` and `kernel`. Our SVM configuration specifically utilized the Radial Basis Function (RBF) kernel to effectively manage the non-linear characteristics of EEG data by mapping the inputs into a higher-dimensional space where a linear decision boundary is feasible based on [29, 52]. Additionally, we tested the kernels *poly* and *sigmoid*. We varied the regularization parameter `C` using 0.025, 0.5, 1.0, 2.0 and 5.0. Rohani and Puthusserypady [52] found 1.0 to be beneficial. We set the maximum amount of iterations to 1,00,000 to allow for adequate training but terminate the process when convergence is unlikely in time. In a grid search, this is important as SVMs might *never* terminate.

4.3 Random Forest

We implemented the RF classifier in Python Scikit-learn toolbox [46]. We performed a grid search on `max_depth`, `max_features`, and `n_estimators`. The training and testing sets were configured in alignment with the methodology adopted for LDA, which involved a shuffled and stratified data distribution. However, unlike LDA, the normalization of training data was intentionally omitted in the RF pipeline. This decision was based on the characteristic performance of RF, which generally shows better accuracy and robustness with non-normalized data due to its inherent handling of feature

variability. The feature vectors used in training and testing the RF classifier were derived from the PSD of each electrode group, providing a rich dataset reflective of the underlying neurological patterns. For the configuration of the RF classifier, the initial setup included a maximum tree depth of 40 and 100 estimators. These hyperparameters were selected based on preliminary studies suggesting their effectiveness for datasets of comparable complexity and size [44, 62, 63].

4.4 Neural Network

We implemented the Vanilla Neural Network (NN) via Scikit-learn toolbox [46]. We performed a grid search on the parameter `hidden_layer_sizes`. We set the parameter `early_stopping` to true to speed up the process and to mitigate overfitting. Also, we increased `max_iter` to 1000 from the 200 default iterations. We used *relu* as activation on the hidden layers and *softmax* function on the last layer to support the classification. For all tests, we used the Adam optimizer.

5 Results

The main objective of this work is to investigate the impact of various choices in terms of feature set, model selection, and generalization for attention detection in VR settings. In Section 5.2, we focused on feature selection to determine the most informative features for classification. In Section 5.4, we examined the effects of different time window lengths for training and testing data. Here, we aimed to identify the optimal window size for accurate classification. Next, in Section 5.5, we compared the performance of different classification algorithms, assessed the effectiveness of each algorithm, and made informed decisions regarding model selection.

5.1 Training Quality

We first look at the accuracy of the training and test for each model. The difference is shown in Figure 2. RF has many overfitted models, aka the test performance is worse than the train quality [9]. Also, we see that some models are under-fitted, so the test accuracy is better than the training accuracy. As both are not desired and lead to not generalizable results, we removed under-fitted results (< -10) and over-fitted results (> 30) from the below analyses.

We fitted a linear mixed model (LMER), estimated using REML and the `nloptwrap` optimizer to predict accuracy based on fit type (formula: accuracy ~ fit). The model included feature combination (`comb`) as a random effect (formula: ~ 1|comb). The model's total explanatory power was substantial, with a conditional R^2 of .78, and the part related to the fixed effects alone (marginal R^2) was .12. The model's intercept, corresponding to `fit = good`, was .70, 95% CI [.66, .74], $t(37195) = 31.35$, $p < .001$. In this model, the effect of `underfitting` was statistically significant and negative, $\beta = -.10$, 95% CI [-.11, -.10], $t(37195) = -135.15$, $p < .001$; standardized $\beta = -.68$, 95% CI [-.69, -.67]. Meanwhile, the effect of `overfitting` was statistically significant and positive, $\beta = .29$, 95% CI [.27, .31], $t(37195) = 26.17$, $p < .001$; standardized $\beta = 1.91$, 95% CI [1.77, 2.06].

These results indicate that models classified as "generalizable" achieved the highest accuracy on average, whereas `underfitting` models were associated with significantly lower accuracy, and `overfitting` models led to moderately higher accuracy at the potential cost of generalizability.

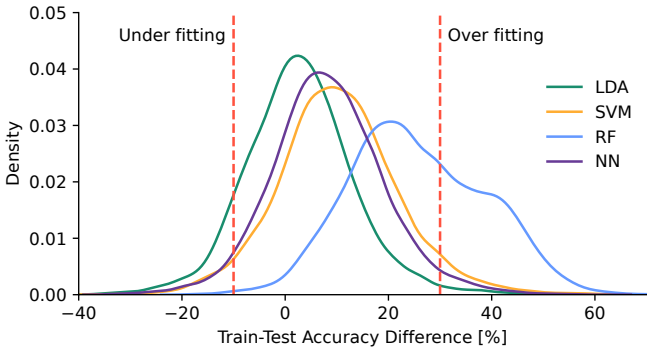


Figure 2: The histograms for the quality of the four models.

5.2 Feature Selection

Next, we look at the features with the overall highest probability of success; see Figure 3. Here, we can identify that the 8 best-performing models consistently rely on the same 8 feature sets[1]. In this context, best-performing models are defined as those achieving, on average, an accuracy of over 80%. We will focus on these models in the remainder of the analyses. Here, we analyze the accuracy across models on the best feature sets by fitting an LMER for each model type. This approach allows us to assess how each model type performs when utilizing the top-performing feature sets. Complete statistical results are available in OSF at https://osf.io/vm9bd/.

5.2.1 LDA. The LMM analysis shows that the model's total explanatory power remains low, with a multiple R^2 of 0.042 (adjusted $R^2 = 0.039$), indicating that approximately 4.2% of the variance in accuracy is explained by the feature combinations. The model's intercept, representing the baseline accuracy level for the reference combination, is 0.837, with an estimated standard error of 0.00398, $t(2197) = 210.38$, $p < 2 \times 10^{-16}$. Several feature combinations exhibit statistically significant positive effects on accuracy. The combination `A-T-B-D-G` has a significant positive impact, with $\beta = 0.0267$, standard error = 0.00549, $t(2197) = 4.87$, $p = 1.23 \times 10^{-6}$. Similarly, the combination `A-T-B-G` shows a significant positive effect, with $\beta = 0.0304$, standard error = 0.00551, $t(2197) = 5.51$, $p = 3.96 \times 10^{-8}$. Another combination, `T-B-D-G`, also yields a statistically significant positive effect, with $\beta = 0.0313$, standard error = 0.00547, $t(2197) = 5.72$, $p = 1.20 \times 10^{-8}$. Lastly, the combination `T-B-G` shows a significant positive impact as well, with $\beta = 0.0238$, standard error = 0.00561, $t(2197) = 4.24$, $p = 2.31 \times 10^{-5}$. In contrast, other combinations, such as `A-T-B-D` and `T-B`, did not emerge as statistically significant predictors of accuracy. These results indicate that certain feature combinations, specifically `A-T-B-D-G`, `A-T-B-G`, `T-B-D-G`, and `T-B-G`, are associated with significant positive effects on accuracy for the LDA model.

5.2.2 SVM. The fitted LMM for SVM showed low explanatory power, with a multiple R^2 of 0.021 (adjusted $R^2 = 0.018$), indicating

[1]Best feature sets: A-T-B, A-T-B-D, A-T-B-G, T-B, T-B-G, T-B-D-G, T-B-D, and A-T-B-D-G

that approximately 2.1% of the variance in accuracy was explained by the feature combinations. The model's intercept, representing the baseline accuracy for the reference combination, was 0.823, with an estimated standard error of 0.004, $t(2314) = 205.01$, $p < 2{\times}10^{-16}$. Within this model, several feature combinations were significant predictors of accuracy. The combination `comb = A-T-B-G` had a statistically significant positive effect, with $\beta = 0.0119$, 95% CI [0.0009, 0.0229], $t(2314) = 2.10$, $p = 0.036$. Similarly, the combination `comb = T-B-D` showed a significant negative effect, with $\beta = -0.0126$, 95% CI [-0.0239, -0.0014], $t(2314) = -2.22$, $p = 0.026$. The combination `comb = T-B-D-G` also had a statistically significant positive effect, with $\beta = 0.0147$, 95% CI [0.0036, 0.0259], $t(2314) = 2.60$, $p = 0.009$. Lastly, the combination `comb = T-B-G` was also a significant positive predictor, with $\beta = 0.0145$, 95% CI [0.0033, 0.0257], $t(2314) = 2.55$, $p = 0.011$. These results suggest that, for the SVM model, specific feature combinations, particularly `A-T-B-G`, `T-B-D-G`, and `T-B-G`, were associated with significantly higher accuracy, while `T-B-D` was associated with a slight reduction in accuracy.

5.2.3 RF. The fitted linear mixed-effects model (LMER) for the RF model demonstrated low explanatory power, with a multiple R^2 of 0.028 (adjusted $R^2 = 0.025$), indicating that approximately 2.8% of the variance in accuracy was explained by the feature combinations. The model's intercept, representing the baseline accuracy for the reference combination, was 0.810, with an estimated standard error of 0.0039, $t(2192) = 210.32$, $p < 2 \times 10^{-16}$. Within this model, several feature combinations were significant predictors of accuracy. The combination `A-T-B-G` had a statistically significant positive effect, with $\beta = 0.0195$, 95% CI [0.0088, 0.0303], $t(2192) = 3.58$, $p = 0.0004$. Conversely, the combination `T-B-D` showed a significant negative effect on accuracy, with $\beta = -0.0148$, 95% CI [-0.0256, -0.0040], $t(2192) = -2.70$, $p = 0.007$. The combination `T-B-D-G` also exhibited a significant positive effect, with $\beta = 0.0108$, 95% CI [0.0001, 0.0215], $t(2192) = 1.98$, $p = 0.048$. Lastly, `T-B-G` was another significant positive predictor, with $\beta = 0.0164$, 95% CI [0.0066, 0.0262], $t(2192) = 3.00$, $p = 0.003$. These results suggest that, for the RF model, specific feature combinations, particularly `A-T-B-G`, `T-B-D-G`, and `T-B-G`, were associated with significantly higher accuracy, while `T-B-D` was associated with a decrease in accuracy.

5.2.4 NN. The LMM for the NN model showed low explanatory power, with a multiple R^2 of X (adjusted $R^2 = Y$), indicating that approximately X% of the variance in accuracy was explained by the feature combinations (note: replace X and Y with actual values if needed). The model's intercept, representing the baseline accuracy for the reference combination, was 0.816, with an estimated standard error of 0.0048, $t(2314) = 168.62$, $p < 2 \times 10^{-16}$. Within this model, several feature combinations were significant predictors of accuracy. The combination `A-T-B-D-G` had a statistically significant positive effect, with $\beta = 0.0171$, 95% CI [0.0040, 0.0302], $t(2314) = 2.54$, $p = 0.011$. Similarly, `A-T-B-G` showed a positive effect on accuracy, with $\beta = 0.0181$, 95% CI [0.0050, 0.0313], $t(2314) = 2.68$, $p = 0.007$. Another significant positive effect was found for the combination `T-B-D-G`, with $\beta = 0.0201$, 95% CI [0.0068, 0.0334], $t(2314) = 2.97$, $p = 0.003$. Lastly, the combination `T-B-G` also demonstrated a significant positive impact, with $\beta = 0.0149$, 95% CI [0.0016, 0.0283], $t(2314) = 2.20$, $p = 0.028$. These results suggest that, for the NN model, feature combinations `A-T-B-D-G`, `A-T-B-G`, `T-B-D-G`, and `T-B-G` were associated with significantly higher accuracy.

5.3 Window Evaluation with Bayesian Analysis

Bayesian data analysis offers several advantages over classical statistical methods, especially in studies with complex hierarchical data structures. Following recent research [41, 51, 58], Bayesian approaches enable us to incorporate prior knowledge, handle small sample sizes, and estimate effect sizes with a given level of precision. Particularly useful in contexts with limited sample sizes and nested data structures, Bayesian methods allow us to quantify the size and uncertainty of effects [31], offering insights into both the presence and absence of effects. We implemented Bayesian linear mixed-effects models to evaluate time window accuracy across models, using the `brms` [6, 8]. Our model choices included four chains with 4,000 iterations and a warm-up of 1,000 samples, ensuring stable posterior estimates and convergence. Priors were weakly informative, with normal priors centered at zero for both the fixed effects and intercept, allowing for flexibility without imposing strong assumptions. The model also included an exponential prior on the random effect standard deviation, tailored to capture variability across combinations. We evaluated model convergence through trace plots, Gelman-Rubin statistics ($\hat{R} < 1.1$), and effective sample sizes, ensuring the reliability of the posterior estimates. Posterior predictive checks were also conducted to assess the model's fit to the observed data visually. These modeling choices, in combination with approximate leave-one-out cross-validation (LOOCV) for model comparison [57], maximize the robustness of our inferences, supporting our aim to identify the optimal time window and assess model performance.

5.3.1 LDA. The Bayesian mixed-effects model assessing the impact of window length on LDA accuracy revealed that increasing the window length consistently improves accuracy. The model included window length as a fixed effect and feature combination as a random effect. The intercept, representing the baseline accuracy for a window length of 1 second, was estimated at 0.79 (95% CI [0.78, 0.81]), with a standard error of 0.01. This result indicates a high baseline accuracy even at the shortest window length. As window length increased, the model showed incremental positive effects on accuracy: a window length of 2 seconds produced an effect of $\beta = 0.04$ (95% CI [0.02, 0.06]), while lengths of 3 and 4 seconds resulted in gains of $\beta = 0.04$ (95% CI [0.02, 0.05]) and $\beta = 0.05$ (95% CI [0.04, 0.07]), respectively. Longer windows continued to enhance accuracy, reaching $\beta = 0.07$ (95% CI [0.05, 0.08]) at lengths of 9, 11, and 13 seconds, suggesting that larger window lengths generally support improved classification performance. The random effect for feature combinations, with an estimated standard deviation of 0.02 (95% CI [0.01, 0.04]), captured minor variability in accuracy across different feature sets. The model's residual error ($\sigma = 0.06$, 95% CI [0.06, 0.06]) indicates consistency in the model's performance.

5.3.2 SVM. The intercept, representing the baseline accuracy for a window length of 1 second, was estimated at 0.77 (95% CI [0.76, 0.79]), with a standard error of 0.01, indicating a reliable starting

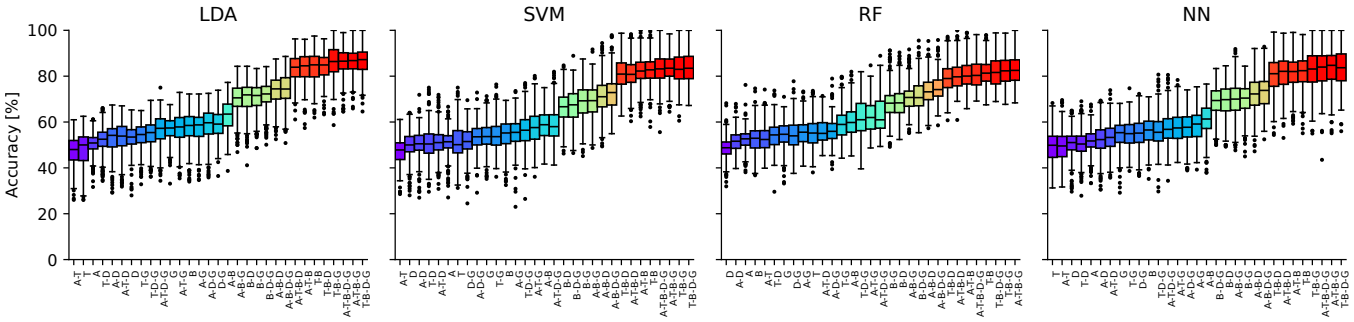


Figure 3: Feature Selection Results.Classification accuracy across various EEG feature combinations for four machine learning models: LDA, SVM, RF, and NN. Each plot displays the distribution of accuracy percentages across different feature sets, alpha (A), theta (T), beta (B), delta (D), gamma (G) — consistently provides the highest accuracy across models

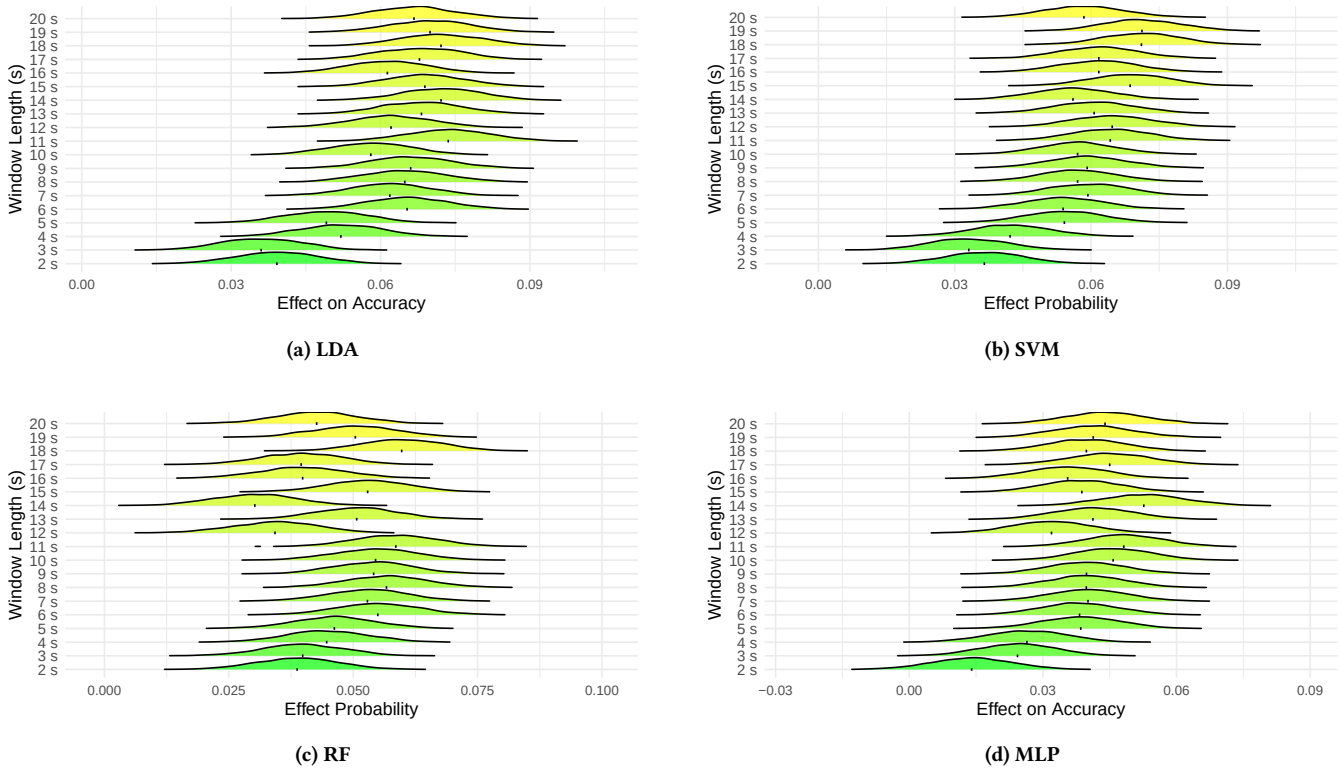
accuracy even at the shortest window length. With additional window length, accuracy showed consistent positive effects. A window length of 2 seconds yielded an increase of $\beta = 0.04$ (95% CI [0.02, 0.05]), while 3 seconds provided a boost of $\beta = 0.03$ (95% CI [0.02, 0.05]). Lengths of 4 and 5 seconds continued this upward trend, with increases of $\beta = 0.04$ (95% CI [0.02, 0.06]) and $\beta = 0.05$ (95% CI [0.04, 0.07]), respectively. This pattern persisted, with window lengths of 7–20 seconds contributing incremental increases up to $\beta = 0.07$ (95% CI [0.05, 0.09]) at various time points, indicating that longer windows generally enhance accuracy for SVM in these settings. The random effect for feature combinations, with an estimated standard deviation of 0.01 (95% CI [0.01, 0.02]), captured minimal variability in accuracy across different feature sets. The residual error for the model ($\sigma = 0.07$, 95% CI [0.06, 0.07]) reflects consistent performance across observations. These results suggest that increasing window length enhances accuracy for SVM, with accuracy improvements plateauing around longer durations.

5.3.3 RF. The Bayesian mixed-effects model examining the effect of window length on RF accuracy demonstrated incremental gains in accuracy as window length increased. In this model, window length was included as a fixed effect, with feature combination treated as a random effect to account for variability across feature sets. The intercept, representing the baseline accuracy at a window length of 1 second, was estimated at 0.77 (95% CI [0.75, 0.78]), with a standard error of 0.01. This high baseline suggests reliable accuracy even with minimal window length. As window length increased, the model showed consistent positive effects on accuracy. Specifically, a window length of 2 seconds resulted in an increase of $\beta = 0.04$ (95% CI [0.02, 0.06]), and window lengths of 3 and 4 seconds showed gains of $\beta = 0.04$ (95% CI [0.02, 0.06]) and $\beta = 0.04$ (95% CI [0.03, 0.06]), respectively. Further increments continued to yield small but steady improvements, with a length of 8 seconds reaching $\beta = 0.06$ (95% CI [0.04, 0.07]) and length 18 showing the highest effect of $\beta = 0.06$ (95% CI [0.04, 0.08]). The random effect for feature combinations, with a standard deviation estimate of 0.01 (95% CI [0.01, 0.03]), indicated minor variability in accuracy across different feature sets, suggesting that the model was relatively stable across combinations. The residual error ($\sigma = 0.06$, 95% CI [0.06, 0.07]) reflects the consistency of the model's performance.

5.3.4 NN. The Bayesian mixed-effects model assessing the influence of window length on NN accuracy revealed consistent accuracy improvements with longer window lengths. The model included window length as a fixed effect and feature combination as a random effect to account for variation across feature sets. The intercept, representing the baseline accuracy for a window length of 1 second, was estimated at 0.78 (95% CI [0.76, 0.80]), with a standard error of 0.01. Accuracy demonstrated progressive increases as window length extended. For instance, a window length of 2 seconds showed an increase of $\beta = 0.04$ (95% CI [0.02, 0.06]), while 3 seconds yielded a gain of $\beta = 0.06$ (95% CI [0.04, 0.08]). Similar gains were observed at 4 and 5 seconds, with estimates of $\beta = 0.05$ (95% CI [0.03, 0.07]) and $\beta = 0.05$ (95% CI [0.03, 0.08]), respectively. By 6 seconds, the accuracy reached $\beta = 0.08$ (95% CI [0.06, 0.10]), indicating notable benefits from longer windows. This pattern held steady across subsequent increments, with estimates remaining positive up to a window length of 20 seconds ($\beta = 0.03$, 95% CI [0.01, 0.05]). The random effect for feature combinations, with an estimated standard deviation of 0.01 (95% CI [0.01, 0.03]), indicated minor variability in accuracy across different feature sets. The residual error for the model ($\sigma = 0.08$, 95% CI [0.08, 0.08]) reflects stable performance across observations.

5.4 Determining Optimal Time Window Lengths via Piecewise Regression

Given the results of our Bayesian linear mixed-effects models, which showed incremental increases in accuracy with length across all models (LDA, SVM, RF, and NN), we have chosen to apply piecewise regression to identify any potential breakpoint where the effect of length on accuracy may stabilize [43]. Piecewise regression is particularly suited for detecting changes in gradient within datasets and can reveal if further increases in time window length yield diminishing returns on accuracy [43]. This approach is particularly effective for identifying structural changes within intervals of the data, allowing the model to accurately capture shifts in the relationship between time window length and accuracy [32, 47]. We aim to identify an optimal time window length that maximizes performance efficiency while maintaining model reliability and reducing unnecessary computational costs. Results are depicted in Figure 6.



(a) LDA

(b) SVM

(c) RF

(d) MLP

Figure 4: Posterior Densities for Window Length Effects on Accuracy Across Different Models. This figure illustrates the effect of varying window lengths on the accuracy of four machine learning models: (a) LDA, (b) SVM, (c) RF, and (d) MLP. Each ridge plot represents the posterior distribution of the effect of window length on accuracy, with colors ranging from green (lower window lengths) to yellow (higher window lengths). The x-axis shows the effect size on accuracy, while the y-axis represents window lengths in seconds. Dashed vertical lines within each ridge indicate the mean effect for that window length. Bayesian analysis shows that increasing window length positively impacts accuracy across all models, although the magnitude of improvement varies. For LDA, accuracy improves steadily, with larger gains up to around 20 seconds. SVM shows consistent accuracy gains with window lengths up to 8 seconds, after which improvements become smaller. RF exhibits stable accuracy growth across all window lengths, with noticeable, gradual gains up to 20 seconds. MLP shows similar incremental improvements, with significant gains up to around 10 seconds before leveling off. These results illustrate that while longer windows generally enhance model accuracy, each model's response to window length varies, indicating different optimal ranges for incremental accuracy gains across models.

5.4.1 LDA Time Window. In the segmented regression analysis for LDA, a significant breakpoint was identified at length = 6.31 seconds (SE = 0.797), as shown in Figure 6. The regression model revealed distinct slopes before and after the breakpoint. For window lengths up to 6.31 seconds, there was a significant positive association between window length and accuracy, $b = 0.0065$, $SE = 0.0015$, $t(15) = 4.33$, $p < 0.001$. After the breakpoint, the slope turned slightly negative, $b = -0.0061$, $SE = 0.0015$, though this change was not statistically significant, $p > 0.05$. The model accounted for a substantial proportion of the variance in accuracy, with $R^2 = 0.83$ and an adjusted $R^2 = 0.80$.

5.4.2 SVM Time Window. In the segmented regression analysis for SVM, a significant breakpoint was identified at length = 6.51 seconds (SE = 0.982), as shown in Figure 6. The model displayed different slopes before and after the breakpoint. Specifically, for window lengths up to 6.51 seconds, there was a significant positive association between window length and accuracy, $b = 0.0056$, $SE = 0.0015$, $t(15) = 3.82$, $p = 0.0017$. Beyond the breakpoint, the slope changed, indicating a slight negative trend, $b = -0.0050$, $SE = 0.0015$, though this change was not statistically significant, $p > 0.05$. The model explained a substantial proportion of the variance in accuracy, with $R^2 = 0.84$ and an adjusted $R^2 = 0.80$.

5.4.3 RF Time Window. In the segmented regression analysis for RF, a breakpoint was identified at length = 6.20 seconds (SE = 1.815), as depicted in Figure 6. The model showed different slopes before and after the breakpoint. For window lengths up to 6.20 seconds, there was a positive but not statistically significant association between window length and accuracy, $b = 0.0039$, $SE = 0.0026$, $t(15) = 1.47$, $p = 0.162$. Beyond the breakpoint, the slope became negative, $b = -0.0046$, $SE = 0.0027$, though this change was not

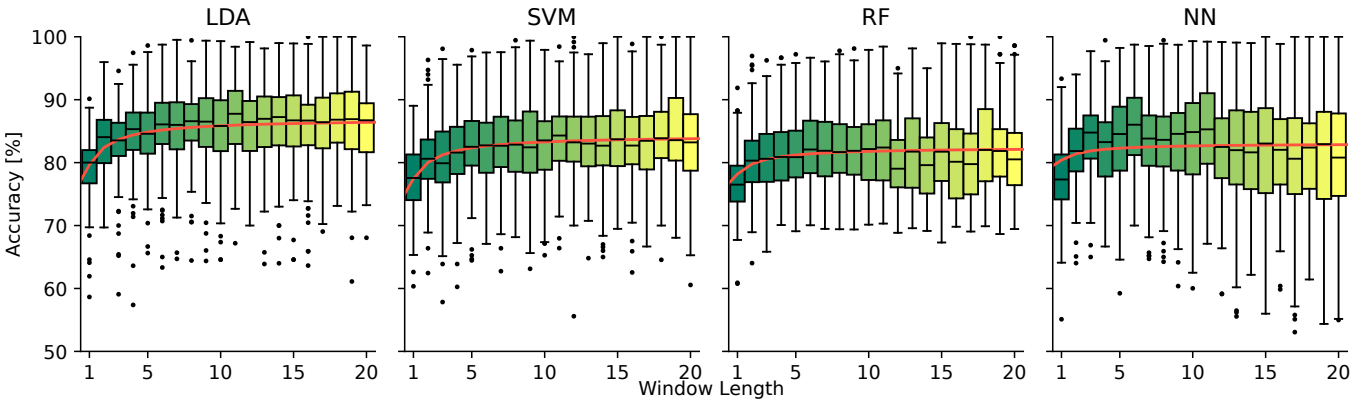


Figure 5: Classification accuracy of LDA, NN, RF, and SVM models as a function of window length. The x-axis represents the window length in seconds, and the y-axis shows the accuracy.

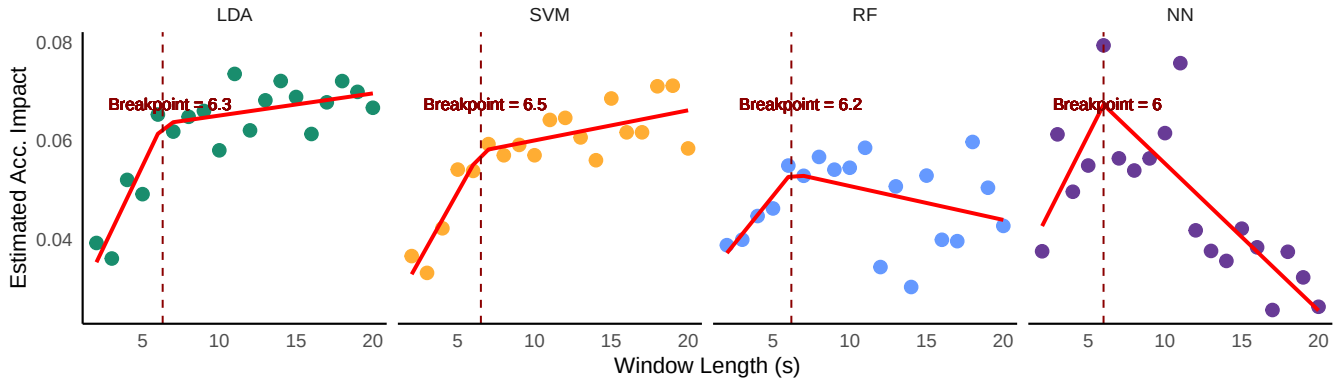


Figure 6: Piecewise regression analysis showing breakpoints in accuracy with increasing window length across four models: LDA, SVM, RF, and NN. The LDA model shows a breakpoint at 6.3s, indicating diminishing returns in accuracy gains beyond this point. The SVM model's breakpoint is at 6.5s, suggesting an early plateau in accuracy. The RF model has a breakpoint at 6.5s. For the NN model, accuracy levels off at a breakpoint of 6s. These breakpoints inform optimal time window lengths for balancing accuracy and efficiency in each model.

statistically significant, $p > 0.05$. The model explained a modest proportion of the variance in accuracy, with $R^2 = 0.23$ and an adjusted $R^2 = 0.07$.

5.4.4 NN Time Window. In the segmented regression analysis for NN, a significant breakpoint was identified at length = 6 seconds (SE = 1.022), as illustrated in Figure 6. The regression model displayed distinct slopes before and after the breakpoint. For window lengths up to 6 seconds, there was a marginally significant positive association between window length and accuracy, $b = 0.0062$, $SE = 0.0031$, $t(15) = 2.00$, $p = 0.064$. Beyond the breakpoint, the slope turned negative, $b = -0.0092$, $SE = 0.0032$, though this change was not statistically significant, $p > 0.05$. The model explained a considerable proportion of the variance in accuracy, with $R^2 = 0.66$ and an adjusted $R^2 = 0.59$.

5.5 Training Split Ratio

We evaluated the effect of *ratio* on accuracy (*acc*) for each model using a linear mixed-effects model, with *length* and *comb* as random effects. Here, we only looked at the eight best-performing feature combinations. The LDA model predicting accuracy from ratio while controlling for length and comb as random effects revealed a significant positive effect of ratio, $\beta = .012$, $SE = .006$, $t(9300) = 2.10$, $p < .05$. The intercept was significant, $\beta = .663$, $SE = .025$, $t(9300) = 26.93$, indicating a baseline accuracy level. For the SVM model, ratio significantly predicted accuracy, $\beta = .020$, $SE = .006$, $t(9300) = 3.38$, $p < .01$, suggesting an increase in accuracy as ratio increased. The intercept was also significant, $\beta = .623$, $SE = 0.024$, $t(9300) = 26.31$. The RF model demonstrated a significant effect of ratio on accuracy, $\beta = .026$, $SE = .005$, $t(9300) = 5.17$, $p < .001$, showing that accuracy increased with ratio. The intercept was significant, $\beta = 0.614$, $SE = .021$, $t(9300) = 29.15$. The NN model indicated a substantial positive effect of ratio on accuracy,

$\beta = .039$, $SE = .006$, $t(9300) = 6.80$, $p < .001$, with a significant intercept, $\beta = .622$, $SE = .023$, $t(9300) = 27.28$. Results are depicted in Figure 7.

5.6 Hyper Parameter

The following section provides a descriptive analysis of the impact of various hyperparameters on model accuracy. Each hyperparameter setting's effect on accuracy is discussed for each model type, highlighting trends and observations.

5.6.1 *RF Hyperparameters.*

Max Depth. The accuracy shows a relatively consistent distribution across different values of *max depth* (5, 10, 15, 20, and None), with median values clustering around similar ranges. However, higher depth values, such as 20 and "None," exhibit slightly more variance, indicating that while increasing maximum depth can influence accuracy variability, there is no clear linear trend in improving median accuracy.

Max Features. For the *max features* parameter, the settings (1, `log`, `sqrt`, and None) yield comparable median accuracy levels, with the `None` setting showing higher variance and a more dispersed accuracy distribution. This suggests that allowing the model to consider all features (`None`) does not necessarily improve median accuracy but increases performance variability.

Number of Estimators. The *n_estimators* parameter, which controls the number of trees, shows that while accuracy distributions are broadly similar across settings (10, 50, 100, 150), *n_estimators* = 100 achieves a slightly higher median and a narrower interquartile range. This suggests that using 100 estimators might provide a balance between accuracy and stability.

5.6.2 *LDA Hyperparameters.*

Shrinkage. The *shrinkage* parameter, with options `auto` and `None`, shows that `auto` tends to yield a higher median accuracy with less variability compared to `None`. This indicates that automated shrinkage might contribute to more consistent performance outcomes.

Solver. For the *solver* parameter, the choices `svd` and `lsqr` exhibit minimal differences in accuracy distribution, with both solvers achieving similar median values. This suggests that the choice of solver may have limited impact on accuracy in this context.

Number of Components. Only one setting for *n_components* (`None`) is shown, resulting in a high median accuracy with minimal variability. This indicates that this setting may be well-suited for maximizing stability in performance.

5.6.3 *SVM Hyperparameters.*

C (Regularization Parameter). The *C* parameter, which controls regularization strength, shows increasing variance in accuracy distribution as its value increases (0.025, 0.5, 1.0, 2.0, 5.0), particularly at $C = 5.0$. Median accuracy remains fairly stable across settings, suggesting that while regularization affects the spread of accuracy outcomes, the central tendency remains less affected.

Kernel. The *kernel* parameter displays notable differences in accuracy distribution across settings (`poly`, `rbf`, `sigmoid`). The `rbf` kernel achieves the highest median accuracy with the least variability, indicating it may be the most suitable for models that prioritize accuracy and robustness.

5.6.4 *NN Hyperparameters.*

Hidden Layer Sizes. Accuracy distributions for various configurations of *hidden layer sizes* (50, 100, 150–50, 50–50, 100–50) exhibit similar median values, though the `150-50` configuration shows slightly higher variability. This suggests that while hidden layer sizes impact variance in accuracy, they do not significantly alter median performance across configurations.

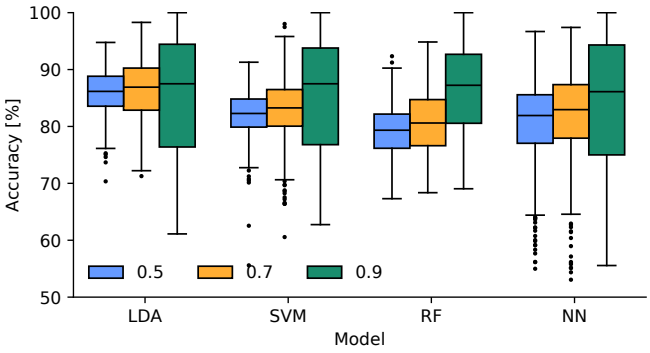


Figure 7: Comparison of classification accuracy across different training split ratios (0.5, 0.7, 0.9) for LDA, SVM, RF, and NN models in a person-independent EEG-based attention detection task. The accuracy generally increases with higher training ratios, with LDA showing the most consistent performance.

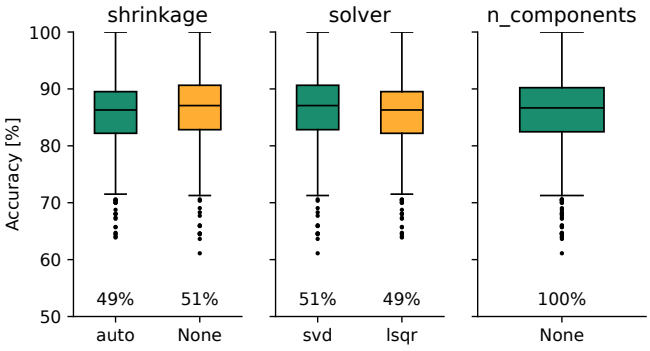
6 Discussion

Understanding the impact of different factors on classification is crucial for developing effective attention-aware systems. We discuss the implications of our findings regarding feature set composition, window length, and training split ratio in adaptive VR.

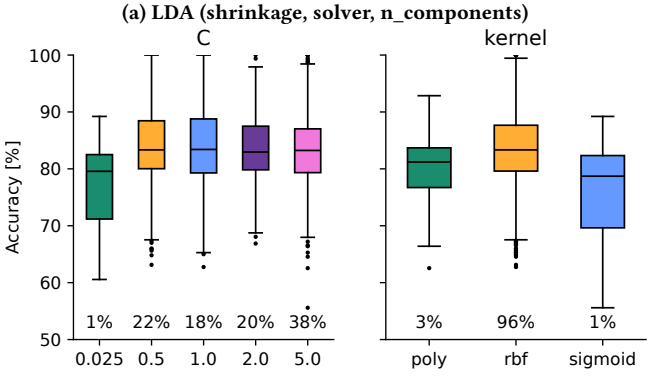
6.1 RQ 1 : Optimal Feature Set

Addressing **RQ1**, we identified eight feature combinations with the highest probability of success across multiple models: A-T-B, A-T-B-D, A-T-B-G, T-B, T-B-G, T-B-D-G, T-B-D, and A-T-B-D-G. These combinations achieved an accuracy of over 80% on average, indicating their robustness for attention classification in VR applications. Similar studies on attention detection in EEG data have emphasized the benefits of multi-band feature sets for improving classifier performance, supporting our approach to using a broad EEG spectrum to enhance attention classification in adaptive systems [4, 11, 38].

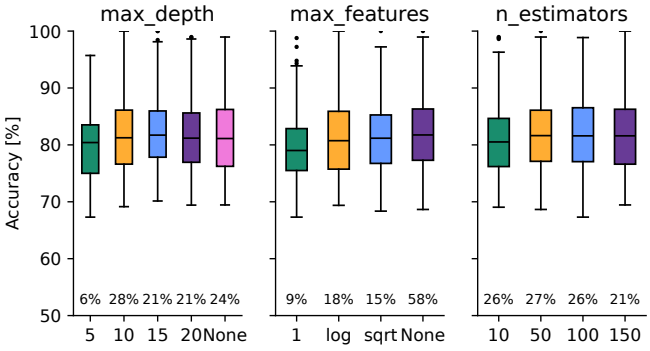
6.1.1 LDA Performance with Specialized Feature Sets. The Linear Discriminant Analysis (LDA) model exhibited nuanced performance across various feature combinations, though the overall explanatory power of the model remained modest, with feature combinations accounting for a small portion of variance in accuracy. Despite this limited explanatory power, specific feature combinations yielded



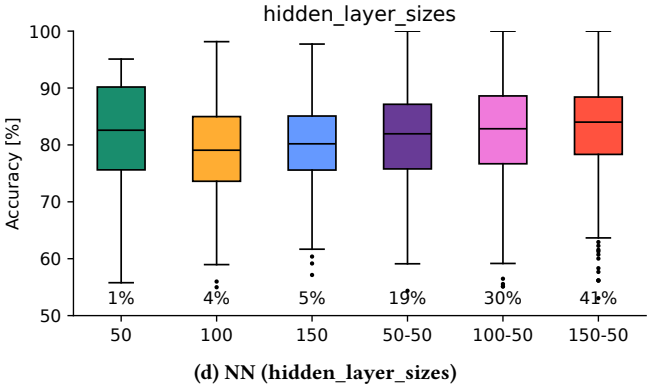
(a) LDA (shrinkage, solver, n_components)



(b) SVM (C, kernel)



(c) RF (max_depth, max_features, n_estimators)



(d) NN (hidden_layer_sizes)

Figure 8: The accuracies of the best models as a result of the grid search. The percentage on the bottom shows how often the grid search returned this configuration parameter as the best model.

significant improvements in classification accuracy, highlight LDA outcome on leverage targeted EEG band sets effectively. Notably, combinations such as A-T-B-D-G, A-T-B-G, T-B-D-G, and T-B-G had a marked positive impact on accuracy. These combinations involve a mix of theta, beta, delta, and gamma bands, suggesting that including delta and gamma bands may provide LDA with richer signal information pertinent to attention classification. The effectiveness of these combinations aligns with previous findings that multi-band EEG data can enhance classifier performance, particularly when the combinations capture complementary neural dynamics [11, 40].

Interestingly, other feature combinations such as A-T-B-D and T-B did not yield significant effects on accuracy, implying that while LDA benefits from broad-spectrum EEG data, certain bands contribute more substantially to its classification capability. Specifically, gamma and delta appear to play pivotal roles in facilitating LDA's performance, as evidenced by their consistent presence in the most effective feature combinations. This finding highlights LDA's adaptability in leveraging distinct EEG frequencies and could inform adaptive VR applications that optimize feature selection based on the task's attention demands.

In summary, the results indicate that although LDA's overall performance remains stable across various combinations, specific sets—particularly those incorporating gamma and delta bands substantially enhance classification accuracy.

6.1.2 *SVM and RF Resilience with Reduced Feature Sets.*

6.1.3 *SVM and RF Resilience with Selective Feature Sets.*

The SVM and RF models demonstrated robustness with reduced feature sets, maintaining significant classification accuracy even when fewer EEG bands were included. This resilience suggests that these models can still provide effective attention classification without requiring all EEG frequency bands, which is advantageous in scenarios where computational efficiency is essential. For SVM, optimal performance was observed with combinations that included the alpha, beta, gamma, and theta bands, particularly in the A-T-B-G, T-B-D-G, and T-B-G combinations, which were associated with significant accuracy improvements. Interestingly, T-B-D exhibited a slight decrease in accuracy, indicating that certain combinations lacking the gamma band may not be as effective for SVM in capturing relevant EEG signals for attention detection.

The RF model similarly displayed selective effectiveness with certain feature combinations, supporting its application in real-time environments where computational resources may be limited. Combinations such as A-T-B-G, T-B-D-G, and T-B-G significantly enhanced accuracy, underscoring the importance of including gamma and theta bands in configurations that optimize RF's performance. On the other hand, the T-B-D combination showed a reduction in accuracy, suggesting that RF may also rely heavily on gamma for improved classification outcomes.

6.1.4 *NN and the Potential for Adaptive Feature Selection.*

The NN model displayed modest performance improvements in response to specific feature configurations, though its accuracy gains were less pronounced than in other models. Notably, the segmented regression analysis revealed a breakpoint at approximately 6 seconds,

suggesting an optimal time window length for maximizing accuracy in attention classification tasks. For window lengths up to this breakpoint, accuracy increased steadily, highlighting that NN models benefit from a balanced window length that captures sufficient temporal information without overfitting to shorter signal fluctuations. Beyond this 6-second window, however, accuracy gains plateaued and began to decline, indicating diminishing returns with longer windows. The observed trends imply that while NN may not exhibit dramatic accuracy boosts with particular feature combinations, it could achieve more consistent performance by employing adaptive feature selection. For example, targeted combinations such as `A-T-B` continue to show potential in boosting classification efficacy within specific contexts. This adaptability underscores the NN model's suitability for VR applications that require flexible attention monitoring across varied user interactions and dynamic environments. By dynamically adjusting the feature sets or time windows based on contextual demands, VR systems can harness the NN model's moderate accuracy improvements while conserving computational resources, aligning well with user-centered and context-responsive VR experiences.

6.2 RQ2 : Identification of Window Length

To identify the most effective time window lengths for classifying attention states in VR (**RQ2**), we . By observing how accuracy changed with each increase in window length, we pinpointed where these improvements began to plateau. This approach helps understand the balance between maximizing accuracy and minimizing computational load, which supports real-time VR applications that rely on efficient processing.

6.2.1 Model-Specific Optimal Window Lengths. The segmented regression analysis across models revealed distinct optimal window lengths, offering insights into the point at which accuracy gains became minimal with further increases in window length. These findings underscore the importance of aligning model-specific window lengths with the characteristics of each algorithm to enhance efficiency and maintain high accuracy in attention classification tasks within VR environments.

For LDA, the analysis identified an optimal window length at approximately 6.31 seconds. Up to this breakpoint, there was a positive association between window length and accuracy, suggesting that extending the window length initially improves classification performance. Beyond 6.31 seconds, however, the slope shifted slightly negative, although this change was not statistically significant. This finding indicates that LDA benefits from a moderately extended window length but does not require very long windows, making it suitable for applications requiring reliable attention tracking without incurring excessive computational demands.

The SVM model displayed a similar breakpoint at around 6.51 seconds, after which accuracy gains also began to diminish. This optimal window length reflects SVM's capacity to maintain high accuracy with moderate amounts of data input. This characteristic is advantageous in VR applications where prompt responsiveness is crucial, as it allows SVM to perform effectively with manageable data windows, balancing accuracy and computational load.

For the RF model, the optimal window length was observed at 6.20 seconds. Although RF's association between window length and accuracy was not statistically significant before or after the breakpoint, the model's performance stabilized around this window length. This rapid stabilization makes RF well-suited for VR applications that prioritize real-time responsiveness and operate under constrained computational resources, as RF can maintain reliable classification performance even with shorter windows.

The NN model demonstrated an optimal window length at approximately 6 seconds. Up to this breakpoint, there was a marginally significant positive association between window length and accuracy, after which accuracy gains turned slightly negative, though not statistically significant. This finding suggests that NN can perform effectively within a relatively short window length, offering a balance between adaptability and efficiency in VR contexts where processing times may vary based on user interactions or application demands.

These model-specific results highlight the optimal window lengths that align with each model's strengths, supporting VR applications in achieving accurate attention detection while optimizing resource use. By calibrating window lengths according to each model's characteristics, VR systems can maximize classification performance and maintain real-time responsiveness in varied attention-monitoring scenarios.

6.2.2 Implications for VR Attention-Aware Adaptive Systems. The consistent breakpoints identified around the 6-second mark across LDA, SVM, RF, and NN models informs for designing EEG-based attention detection in VR applications. This similarity suggests a potential temporal threshold that captures key attention-related EEG patterns, offering a standardized window length that could work effectively across multiple models. Rather than requiring model-specific window durations, VR systems could leverage this shared 6-second breakpoint, simplifying the implementation of attention-aware adaptive systems.

Aligning VR system design with this approximate 6-second window length balances the trade-off between accuracy and computational efficiency. This threshold appears to represent an effective duration for capturing EEG signal information, beyond which additional data yields minimal gains in classification accuracy. For applications needing reliable attention tracking, particularly in resource-limited settings, a standardized window length enables consistent performance across models, reducing the complexity of tailoring unique configurations for each algorithm.

The convergence of breakpoints across models also suggests that EEG signals carry a characteristic attention-related information density within this time frame. This insight supports the hypothesis that EEG data for attention classification may exhibit a natural temporal resolution, beneficial for adaptive VR environments where attention shifts must be detected and responded to promptly. By standardizing around this effective window length, VR systems can achieve a high degree of compatibility across different machine-learning models, supporting robust and efficient real-time adaptations in various VR contexts.

6.3 RQ3 : Training Ratio

We analyzed effect of the training split ratio on the performance of various model to address **RQ3**. Generally, increasing the proportion of training data tends to enhance model performance. This trend

identifies the benefit of extensive training datasets in improving accuracy, a finding that aligns with expectations across machine learning applications [7, 39].

For LDA, we observed that accuracy remained relatively stable as the training data ratio increased, suggesting robustness against variations in training data volume. In contrast, the impact on other models was more complex. Specifically, the NN and RF models showed improved performance with larger training datasets. This improvement likely reflects their greater capacity for capturing and generalizing complex patterns from more extensive data.

However, the performance of the Support Vector Machine (SVM) model presents a unique case. While performance improves with initial increases in the training ratio, there appears to be a threshold beyond which additional data does not contribute to better outcomes and may even reduce performance. This phenomenon should not be immediately labeled as overfitting, which typically relates to overly complex models for the available data, but rather as a potential inefficiency in learning from excessively large datasets.

These findings suggest that training data allocation should be tailored to the model used. Systems employing NN or RF could benefit from larger datasets to fully utilize their pattern recognition capabilities. Conversely, it is important to find an optimal training data limit for SVMs to avoid unnecessary computational expenses and potential performance declines.

6.4 Recommendations for Hyperparameter Selection

Based on our findings, we recommend specific hyperparameter settings for each model to optimize accuracy and consistency in EEG-based attention detection in VR environments.

6.4.1 LDA. Using "auto" for shrinkage improves stability by adjusting shrinkage based on the data, leading to more reliable results. Solver choice ("svd" or "lsqr") does not significantly affect accuracy, so either option is fine. Keeping `n_components` at None maintains high accuracy by preserving all relevant features, which enhances performance stability.

6.4.2 RF. A max depth of 10–15 is optimal, as it captures data patterns without excessive variability, which can arise with deeper trees. Using "sqrt" for max features is ideal because it provides consistent accuracy while controlling for unnecessary complexity. Setting `n_estimators` to 100 balances accuracy with computational efficiency, as adding more trees yields diminishing returns.

6.4.3 SVM. A moderate `C` value around 1.0 provides strong accuracy without overfitting. The RBF kernel is the best choice, as it captures nonlinear EEG patterns more effectively, making the model more robust and adaptable for VR applications.

6.4.4 NN. A single hidden layer with 100 units achieves a good balance between accuracy and computational demand. This configuration minimizes complexity while maintaining stable performance and is suitable for real-time applications.

6.5 Limitations & Future Work

We acknowledge several limitations in our study, including feature generation, task generalization, and the constraints due to dataset size. These limitations highlight areas for future investigations to improve online attention detection within VR.

Concerning feature generation, our study utilized an EEG setup with 64 electrodes. However, future research could explore the feasibility of reducing the number of electrodes to enhance BCI usability and minimize their intrusiveness. Considering the trade-off between the number of electrodes and classification accuracy is crucial. While increasing the number of electrodes might improve accuracy, it could also lead to increased system complexity and reduced user comfort. Thus, achieving an optimal balance between electrode amount, accuracy, and real-time performance requires meticulous evaluation in various usage contexts.

In future work, we plan to implement the online performance of our classifiers for attention detection. Leveraging the reccomendations from our model selection, feature sets, and time windows developed during offline classification will guide the transition to online applications. A significant challenge here is the real-time processing and analysis of EEG data, which necessitates the development of optimized and lightweight algorithms capable of handling streaming EEG data effectively. For online signal processing, approaches such as Common Spatial Patterns (CSP) and Spatio-spectral filters are valuable. CSP enhances the discriminative power of EEG signals by identifying spatial filters that highlight relevant brain activity patterns [5]. Spatio-spectral filters integrate spatial and spectral information to refine EEG signal representation and emphasize frequency-specific modulations [36]. These techniques must be adapted for real-time processing, which involves optimizing algorithms and implementing parallel processing to manage the computational demands.

Additionally, we aim to generalize our developed models to other EEG datasets, such as those used in AR settings [60] or through the entire Mixed Reality continuum [10] for both internal and external attention detection. Building on the previous work [62], we plan to leverage their findings to enhance the transferability of our models and assess their applicability to similar tasks. Given the primary distinction in EEG data for both attentional directions is the focus of attention, we anticipate that the results of our study will be replicable in other internal/external attention tasks. Nonetheless, evaluating the task dependency of the results is critical to confirm their generalizability.

Finally, we propose the exploration of Generative Adversarial Networks (GANs) to aid the generalization of our models to different tasks. GANs have proven successful in generating artificial data that closely mimic real-world samples, including recent advancements in artificial EEG data [21, 28]. Using GANs, we can generate additional data samples not present in the original dataset, facilitating data augmentation and expanding the training data with unseen examples. This approach will enable improved model decoding and generalization capabilities across different attention tasks.

7 Conclusion

This work systematically evaluated model accuracy, EEG features, window segmentation, and person-independent classification for future online use in VR attention detection. Inspired by previous AR research [62], we compared different machine learning models,

including LDA, SVM, Random Forest, and Neural Net. Our work contributes to the understanding and design of attention detection in VR environments by addressing different features, models, and data amounts for classification. It provides insights into the optimal feature set, the impact of window segmentation, and the training performance of different machine learning models.

8 Open Science

Analysis scripts and models can be accessed on the Open Science Framework via https://osf.io/vm9bd/.

Acknowledgments

Francesco Chiossi and Sven Mayer were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with Project ID 251654672 TRR 161.

References

[1] Allison P Anderson, Michael D Mayer, Abigail M Fellows, Devin R Cowan, Mark T Hegel, and Jay C Buckey. 2017. Relaxation with immersive natural scenes presented using virtual reality. *Aerospace medicine and human performance* 88, 6 (2017), 520–526. https://doi.org/10.3357/AMHP.4747.2017

[2] Aurélien Appriou, Andrzej Cichocki, and Fabien Lotte. 2018. Towards Robust Neuroadaptive HCI: Exploring Modern Machine Learning Methods to Estimate Mental Workload From EEG Signals. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188617

[3] Christopher Baker and Stephen H Fairclough. 2022. Adaptive virtual reality. In *Current Research in Neuroadaptive Technology*. Elsevier, Amsterdam, Netherlands, 159–176. https://doi.org/10.1016/B978-0-12-821413-8.00014-2

[4] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. 2015. Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448* (2015). https://doi.org/10.48550/arXiv.1511.06448

[5] Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-robert Muller. 2008. Optimizing Spatial filters for Robust EEG Single-Trial Analysis. *IEEE Signal Processing Magazine* 25, 1 (2008), 41–56. https://doi.org/10.1109/MSP.2008.4408441

[6] Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software* 80 (2017), 1–28.

[7] Henry Candra, Mitchell Yuwono, Rifai Chai, Ardi Handojoseno, Irraivan Elamvazuthi, Hung T Nguyen, and Steven Su. 2015. Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine. In *2015 37th Annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, New York, NY, USA, 7250–7253.

[8] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 76 (2017), 1. https://doi.org/10.18637/jss.v076.i01

[9] Gavin C Cawley and Nicola LC Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* 11 (2010), 2079–2107.

[10] Francesco Chiossi, Yassmine El Khaoudi, Changkun Ou, Ludwig Sidenmark, Abdelrahman Zaky, Tiare Feuchtner, and Sven Mayer. 2024. Evaluating Typing Performance in Different Mixed Reality Manifestations using Physiological Features. *Proc. ACM Hum.-Comput. Interact.* 8, ISS, Article 542 (Oct. 2024), 30 pages. https://doi.org/10.1145/3698142

[11] Francesco Chiossi, Changkun Ou, Carolina Gerhardt, Felix Putze, and Sven Mayer. 2023. Designing and Evaluating an Adaptive Virtual Reality System using EEG Frequencies to Balance Internal and External Attention States. *arXiv preprint arXiv:2311.10447* (2023). https://doi.org/10.48550/arXiv.2311.10447

[12] Francesco Chiossi, Yagiz Turgut, Robin Welsch, and Sven Mayer. 2023. Adapting Visual Complexity Based on Electrodermal Activity Improves Working Memory Performance in Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 7, MHCI, Article 196 (sep 2023), 26 pages. https://doi.org/10.1145/3604243

[13] Francesco Chiossi, Robin Welsch, Steeven Villa, Lewis Chuang, and Sven Mayer. 2022. Virtual Reality Adaptation Using Electrodermal Activity to Support the User Experience. *Big Data and Cognitive Computing* 6, 2 (2022), 55. https://doi.org/10.3390/bdcc6020055

[14] Francesco Chiossi, Johannes Zagermann, Jakob Karolus, Nils Rodrigues, Priscilla Balestrucci, Daniel Weiskopf, Benedikt Ehinger, Tiare Feuchtner, Harald Reiterer, Lewis L. Chuang, Marc Ernst, Andreas Bulling, Sven Mayer, and Albrecht Schmidt. 2022. Adapting visualizations and interfaces to the user. *it - Information Technology* (2022). https://doi.org/10.1515/itit-2022-0035

[15] Marvin M Chun, Julie D Golomb, and Nicholas B Turk-Browne. 2011. A taxonomy of external and internal attention. *Annual review of psychology* 62 (2011). https://doi.org/10.1146/annurev.psych.093008.100427

[16] Giorgia Cona, Francesco Chiossi, Silvia Di Tomasso, Giovanni Pellegrino, Francesco Piccione, Patrizia Bisiacchi, and Giorgio Arcara. 2020. Theta and alpha oscillations as signatures of internal and external attention to delayed intentions: A magnetoencephalography (MEG) study. *NeuroImage* 205 (2020), 116295. https://doi.org/10.1016/j.neuroimage.2019.116295

[17] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. 2019. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of neural engineering* 16, 3 (2019), 031001. https://doi.org/10.1088/1741-2552/ab0ab5

[18] Arindam Dey, Jane Phoon, Shuvodeep Saha, Chelsea Dobbins, and Mark Billinghurst. 2020. A Neurophysiological Approach for Measuring Presence in Immersive Virtual Environments. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, New York, NY, USA, 474–485. https://doi.org/10.1109/ISMAR50242.2020.00072

[19] Arindam Dey, Thammathip Piumsomboon, Youngho Lee, and Mark Billinghurst. 2017. Effects of Sharing Physiological States of Players in a Collaborative Virtual Reality Gameplay. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4045–4056. https://doi.org/10.1145/3025453.3026028

[20] Dennis Dietz, Carl Oechsner, Changkun Ou, Francesco Chiossi, Fabio Sarto, Sven Mayer, and Andreas Butz. 2022. Walk This Beam: Impact of Different Balance Assistance Strategies and Height Exposure on Performance and Physiological Arousal in VR. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology* (Tsukuba, Japan) *(VRST '22)*. Association for Computing Machinery, New York, NY, USA, Article 32, 12 pages. https://doi.org/10.1145/3562939.3567818

[21] Hendrik Eilts and Felix Putze. 2022. Is that real? A multifaceted evaluation of the quality of simulated EEG signals for passive BCI. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, New York, NY, USA, 2639–2644. https://doi.org/10.1109/SMC53654.2022.9945093

[22] Kate C Ewing, Stephen H Fairclough, and Kiel Gilleade. 2016. Evaluation of an adaptive game that uses EEG measures validated during the design process as inputs to a biocybernetic loop. *Frontiers in human neuroscience* 10 (2016), 223. https://doi.org/10.3389/fnhum.2016.00223

[23] Stephen H Fairclough. 2009. Fundamentals of physiological computing. *Interacting with computers* 21, 1-2 (2009), 133–145. https://doi.org/10.1016/j.intcom.2008.10.011.

[24] Dongyu Gong and Jan Theeuwes. 2021. A saliency-specific and dimension-independent mechanism of distractor suppression. *Attention, Perception, & Psychophysics* 83 (2021), 292–307. https://doi.org/10.3758/s13414-020-02142-8

[25] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, and Lauri Parkkonen. 2013. MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience* 7 (2013), 267. https://doi.org/10.3389/fnins.2013.00267

[26] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. 2023. Beyond Self-Attention: External Attention Using Two Linear Layers for Visual Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 5 (2023), 5436–5447. https://doi.org/10.1109/TPAMI.2022.3211006

[27] Thalía Harmony, Thalía Fernández, Juan Silva, Jorge Bernal, Lourdes Díaz-Comas, Alfonso Reyes, Erzsébet Marosi, Mario Rodríguez, and Miguel Rodríguez. 1996. EEG delta activity: an indicator of attention to internal processing during performance of mental tasks. *International journal of psychophysiology* 24, 1-2 (1996), 161–171. https://doi.org/10.1016/S0167-8760(96)00053-0

[28] Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. 2018. EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. *arXiv preprint arXiv:1806.01875* (2018).

[29] Jupitara Hazarika, Piyush Kant, Rajdeep Dasgupta, and Shahedul Haque Laskar. 2018. Neural modulation in action video game players during inhibitory control function: an EEG study using discrete wavelet transform. *Biomedical Signal Processing and Control* 45 (2018), 144–150.

[30] Kenji Katahira, Yoichi Yamazaki, Chiaki Yamaoka, Hiroaki Ozaki, Sayaka Nakagawa, and Noriko Nagata. 2018. EEG correlates of the flow state: A combination of increased frontal theta and moderate frontocentral alpha rhythm in the mental arithmetic task. *Frontiers in psychology* 9 (2018), 300. https://doi.org/10.3389/fpsyg.2018.00300

[31] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4521–4532. https://doi.org/10.1145/2858036.2858465

[32] Taehyeong Kim, Hyungu Lee, and Hayoung Choi. 2024. Improved identification of breakpoints in piecewise regression and its applications. *arXiv preprint arXiv:2408.13751* (2024).

[33] Anastasia Kiyonaga and Tobias Egner. 2013. Working memory as internal attention: Toward an integrative account of internal and external selection processes. *Psychonomic bulletin & review* 20 (2013), 228–242. https://doi.org/10.3758/s13423-012-0359-y

[34] Christian A. Kothe and Scott Makeig. 2011. Estimation of task workload from EEG data: New and current tools and perspectives. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, New York, NY, USA, 6547–6551. https://doi.org/10.1109/IEMBS.2011.6091615

[35] Laurens R Krol and Thorsten O Zander. 2022. Defining neuroadaptive technology: the trouble with implicit human-computer interaction. In *Current Research in Neuroadaptive Technology*. Elsevier, Amsterdam, Netherlands, 17–42. https://doi.org/10.1016/B978-0-12-821413-8.00007-5

[36] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller. 2005. Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Transactions on Biomedical Engineering* 52, 9 (2005), 1541–1548. https://doi.org/10.1109/TBME.2005.851521

[37] Seokbeen Lim, Mina Yeo, and Gilwon Yoon. 2019. Comparison between concentration and immersion based on EEG analysis. *Sensors* 19, 7 (2019), 1669. https://doi.org/10.3390/s19071669

[38] Xingyu Long, Sven Mayer, and Francesco Chiossi. 2024. Multimodal Detection of External and Internal Attention in Virtual Reality using EEG and Eye Tracking Features. In *Proceedings of Mensch Und Computer 2024* (Karlsruhe, Germany) *(MuC '24)*. Association for Computing Machinery, New York, NY, USA, 29–43. https://doi.org/10.1145/3670653.3670657

[39] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences* 250 (2013), 113–141.

[40] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. 2018. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering* 15, 3 (2018), 031005.

[41] Tianwen Ma, Yang Li, Jane E Huggins, Ji Zhu, and Jian Kang. 2022. Bayesian inferences on neural activity in EEG-based brain-computer interface. *J. Amer. Statist. Assoc.* 117, 539 (2022), 1122–1133.

[42] Elisa Magosso, Francesca De Crescenzio, Giulia Ricci, Sergio Piastra, and Mauro Ursino. 2019. EEG alpha power is modulated by attentional changes during cognitive tasks and virtual reality immersion. *Comp. Intelligence and Neuroscience* (2019). https://doi.org/10.1155/2019/7051079

[43] Victor E McZgee and Willard T Carleton. 1970. Piecewise regression. *J. Amer. Statist. Assoc.* 65, 331 (1970), 1109–1124.

[44] Zainab Mohamed, Mohamed El Halaby, Tamer Said, Doaa Shawky, and Ashraf Badawi. 2018. Characterizing focused attention and working memory using EEG. *Sensors* 18, 11 (2018), 3743.

[45] Rebecca Patient, Fawaz Ghali, Hoshang Kolivand, William Hurst, and Nigel John. 2021. Application of Virtual Reality and Electrodermal Activity for the Detection of Cognitive Impairments. In *14th International Conference on Developments in eSystems Engineering (DeSE)*. IEEE, New York, NY, USA, 156–161. https://doi.org/10.1109/DeSE54285.2021.9719442

[46] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[47] Charlie Pilgrim. 2021. piecewise-regression (aka segmented regression) in Python. *Journal of Open Source Software* 6, 68 (2021).

[48] Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. 2019. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage* 198 (2019), 181–197. https://doi.org/10.21105/joss.04484

[49] Felix Putze, Maximilian Scherer, and Tanja Schultz. 2016. Starring into the Void? Classifying Internal vs. External Attention from EEG. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (Gothenburg, Sweden) *(NordiCHI '16)*. ACM, New York, NY, USA, Article 47, 4 pages. https://doi.org/10.1145/2971485.2971555

[50] Giulia Ricci, Francesca De Crescenzio, Sandhya Santhosh, Elisa Magosso, and Mauro Ursino. 2022. Relationship between electroencephalographic data and comfort perception captured in a Virtual Reality design environment of an aircraft cabin. *Scientific Reports* 12, 1 (2022), 10938.

[51] Constanze Riha, Dominik Güntensperger, Tobias Kleinjung, and Martin Meyer. 2020. Accounting for heterogeneity: mixed-effects models in resting-state EEG data in a sample of tinnitus sufferers. *Brain topography* 33 (2020), 413–424.

[52] Darius A Rohani and Sadasivan Puthusserypady. 2015. BCI inside a virtual reality classroom: a potential training tool for attention. *EPJ Nonlinear Biomedical Physics* 3 (2015), 1–14.

[53] James B Rowe, Ivan Toni, Oliver Josephs, Richard SJ Frackowiak, and Richard E Passingham. 2000. The prefrontal cortex: response selection or maintenance within working memory? *Science* 288, 5471 (2000), 1656–1660. https://doi.org/10.1126/science.288.5471.1656

[54] Hannah J Scheibner, Carsten Bogler, Tobias Gleich, John-Dylan Haynes, and Felix Bermpohl. 2017. Internal and external attention and the default mode network. *Neuroimage* 148 (2017), 381–389. https://doi.org/10.1016/j.neuroimage.2017.01.044

[55] Rhaíra Helena Caetano e Souza and Eduardo Lázaro Martins Naves. 2021. Attention detection in virtual environments using EEG signals: A scoping review. *frontiers in physiology* 12 (2021), 727840. https://doi.org/10.3389/fphys.2021.727840

[56] Sokkeang Try, Kriengsak Panuwatwanich, Ganchai Tanapornraweekit, and Manop Kaewmoracharoen. 2021. Virtual reality application to aid civil engineering laboratory course: A multicriteria comparative study. *Computer Applications in Engineering Education* 29, 6 (2021), 1771–1792. https://doi.org/10.1002/cae.22422

[57] Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing* 27 (2017), 1413–1432.

[58] Antonino Visalli, Maria Montefinese, Giada Viviani, Livio Finos, Antonino Vallesi, and Ettore Ambrosini. 2024. lmeEEG: Mass linear mixed-effects modeling of EEG data with crossed random effects. *Journal of Neuroscience Methods* 401 (2024), 109991.

[59] Francesca Vitali, Cantor Tarperi, Jacopo Cristini, Andrea Rinaldi, Arnaldo Zelli, Fabio Lucidi, Federico Schena, Laura Bortoli, and Claudio Robazza. 2019. Action monitoring through external or internal focus of attention does not impair endurance performance. *Frontiers in Psychology* 10 (2019), 535. https://doi.org/10.3389/fpsyg.2019.00535

[60] Lisa-Marie Vortmann, Felix Kroll, and Felix Putze. 2019. EEG-based classification of internally-and externally-directed attention in an augmented reality paradigm. *Frontiers in human neuroscience* 13 (2019), 348.

[61] Lisa-Marie Vortmann and Felix Putze. 2020. Attention-Aware Brain Computer Interface to Avoid Distractions in Augmented Reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382889

[62] Lisa-Marie Vortmann and Felix Putze. 2021. Exploration of Person-Independent BCIs for Internal and External Attention-Detection in Augmented Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 80 (jun 2021), 27 pages. https://doi.org/10.1145/3463507

[63] Lisa-Marie Vortmann, Pascal Weidenbach, and Felix Putze. 2022. AtAwAR Translate: Attention-Aware Language Translation Application in Augmented Reality for Mobile Phones. *Sensors* 22, 16 (2022), 6160. https://doi.org/10.3390/s22166160

[64] Jue Wang, Nan Yan, Hailong Liu, Mingyu Liu, and Changfeng Tai. 2007. Brain-computer interfaces based on attention and complex mental tasks. In *Digital Human Modeling: First Inter. Conference on Digital Human Modeling*. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-540-73321-8_54