life

MDPI

Article

On the Influence of Aging on Classification Performance in the Visual EEG Oddball Paradigm Using Statistical and Temporal Features

Nina Omejc [1,2,*], Manca Peskar [3,4], Aleksandar Miladinović [5], Voyko Kavcic [6,7], Sašo Džeroski [1] and Uros Marusic [3,8]

1 Department of Knowledge Technologies, Jožef Stefan Institute, 1000 Ljubljana, Slovenia
2 Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia
3 Institute for Kinesiology Research, Science and Research Centre Koper, 6000 Koper, Slovenia
4 Biological Psychology and Neuroergonomics, Department of Psychology and Ergonomics, Faculty V: Mechanical Engineering and Transport Systems, Technische Universität Berlin, 10623 Berlin, Germany
5 Department of Ophthalmology, Institute for Maternal and Child Health-IRCCS Burlo Garofolo, 34137 Trieste, Italy
6 Institute of Gerontology, Wayne State University, Detroit, MI 48202, USA
7 International Institute of Applied Gerontology, 1000 Ljubljana, Slovenia
8 Department of Health Sciences, Alma Mater Europaea—ECM, 2000 Maribor, Slovenia
* Correspondence: nina.omejc@ijs.si

Abstract: The utilization of a non-invasive electroencephalogram (EEG) as an input sensor is a common approach in the field of the brain–computer interfaces (BCI). However, the collected EEG data pose many challenges, one of which may be the age-related variability of event-related potentials (ERPs), which are often used as primary EEG BCI signal features. To assess the potential effects of aging, a sample of 27 young and 43 older healthy individuals participated in a visual oddball study, in which they passively viewed frequent stimuli among randomly occurring rare stimuli while being recorded with a 32-channel EEG set. Two types of EEG datasets were created to train the classifiers, one consisting of amplitude and spectral features in time and another with extracted time-independent statistical ERP features. Among the nine classifiers tested, linear classifiers performed best. Furthermore, we show that classification performance differs between dataset types. When temporal features were used, maximum individuals' performance scores were higher, had lower variance, and were less affected overall by within-class differences such as age. Finally, we found that the effect of aging on classification performance depends on the classifier and its internal feature ranking. Accordingly, performance will differ if the model favors features with large within-class differences. With this in mind, care must be taken in feature extraction and selection to find the correct features and consequently avoid potential age-related performance degradation in practice.

Keywords: aging; EEG; machine learning; classification; BCI; visual oddball study

Citation: Omejc, N.; Peskar, M.; Miladinović, A.; Kavcic, V.; Džeroski, S.; Marusic, U. On the Influence of Aging on Classification Performance in the Visual EEG Oddball Paradigm Using Statistical and Temporal Features. *Life* **2023**, *13*, 391. https://doi.org/10.3390/life13020391

Academic Editors: Alessandra Anzolin, Jorge Bosch-Bayard, Giuseppe Augusto Chiarenza and Rubén Pérez-Elvira

Received: 2 January 2023
Revised: 23 January 2023
Accepted: 28 January 2023
Published: 31 January 2023

1. Introduction

The loss of communication pathways that enable interaction with the environment often has serious consequences for the patient. With ongoing advances and increasing acceptance of robotics and brain-computer interfaces (BCI) in neurorehabilitation, some negative impacts can be mitigated. When it is not otherwise possible, various devices (e.g., a neuroprosthesis, a wheelchair, a speller, or a computer cursor, see [1]) can be controlled by an electroencephalogram (EEG). EEG has many suitable properties—it is a non-invasive, safe, relatively simple, and inexpensive neuroimaging technique with a high temporal resolution that provides extensive information about brain activity. On the other hand, the gathered EEG data are complex, non-linear, non-stationary, high-dimensional signals that

are contaminated with noise and inter-individual differences [2–4]. Consequently, efficient and reliable EEG data analysis is essential [5].

EEG BCIs provide the user with an alternative way of acting on the world. While measuring EEG activity, important features are extracted from the brain signal and analyzed. Then, the obtained information is in real-time translated into specific command, executed by a device that is implementing the user's intentions [6]. Among the most relevant and common EEG features used in BCIs are the event-related potentials (ERPs) [5,7,8]. ERPs are relatively small voltage spikes generated by the underlying cascaded cortical processes and are triggered by external events such as visual stimuli. The most common visual ERPs are the positive P1 and the negative N1 waves, recorded from the occipital electrodes around 100 ms after the stimulus onset [9]. When the facial stimulus is presented, the N1 wave includes the N170 component that occurs at approximately 170 ms and is strongly related to the early perception of faces [10]. Beside early visual components, the P3 wave is an important late endogenous positive component, occurring around 300 ms after the stimulus presentation. The P3, most prominently observed in the parieto-central regions, is classically elicited in an oddball study, where randomly occurring rare stimuli capture attention in between the less relevant, frequent stimuli [11].

An important point we explore further in this study is that a healthy aging brain, ignoring individual variability, exhibits a number of differences in EEG activity compared to the brain of young adults [12–14]. While the basic neural mechanisms are maintained, older adults show an overall reduction in power across all frequency bands, indicative of general slowing [15–17]. Moreover, many studies have shown ERP age-related changes. Although ERPs are highly task-specific, the general consensus is that aging delays the latency of the visual (and other sensory) ERPs [12,18–21], while the effects on the sensory ERP amplitudes are not so clear. Studies mainly report no differences [19–21] or a decrease in amplitude [21–24]. Interestingly, following the compensatory hypothesis of aging [25], it has been shown that amplitudes of the posterior components decrease with age, while the amplitudes of the components in the prefrontal areas increase with age in order to maintain good performance [22]. As neurorehabilitation often includes older patients, possible brain changes must be considered when designing the EEG BCI rehabilitation process using automatic machine learning approaches.

To better understand how age-related differences might affect the BCI closed-loop rehabilitation process, we performed a binary classification task with a visual EEG oddball trial that elicited both early visual ERPs and the P3 wave. When designing a classification task, the classifier and feature selection are the two most important steps. From the plethora of possible classifiers, we decided to test the aging effect with (i) three linear classifiers, namely, Linear Discriminant Analysis (LDA), Support Vector Classifier (SVC) with no kernel, and Logistic Regression (LR), (ii) three nonlinear classifiers, namely Decision Tree Classifier (tree), K-Nearest Neighbours (KNN) and the SVC with radial basis function (RBF) kernel, (iii) as well as three ensemble methods, namely Random Forest (RF), Adaptive Boosting (AdaBoost) and Extreme Gradient Boosting (XGB). The classifiers were chosen as the representatives of each class, as they commonly appeared in the EEG BCI-related literature [1,6,8,26,27] or are classical representatives of specific method type. Especially LDA [28–30] and SVM [31–33] have been highly used in practice, but also KNN [34], LR [35] and RF [36]. Despite the current high interest in deep neural network (DNN) architectures (see e.g., [6,37–42]), these were not considered here due to the limited training data available and the number of hyperparameters that would need to be defined.

When using standard classifiers for high-dimensional EEG data, the feature extraction and selection steps are of great importance. A recent article reviewed 147 EEG features that can be extracted from the EEG signal [27], grouped by features in the time-domain, frequency-domain, time-frequency domain, nonlinear features, entropy, spatio-temporal features, and complex networks features. The review does not imply that certain features are necessarily better than others, but it does suggest that using various feature types can lead to improved classification performance. Following this notion, we created two

types of datasets for training. The first was time-independent and contained relevant ERP statistical parameters, extracted from the time series. For instance, each ERP component can be parameterized by the peak amplitude, the mean amplitude, the peak latency, and the fractional 50% peak latency [43]. The second dataset included time-dependent features, particularly the amplitude and the signal power at a given frequency over time. While too large to be useful in an online application, such an approach enables better investigation of the possible age differences and variations in feature importance over time.

Related Work

As already described, the age-related changes in ERPs properties have been known for a while and many studies have focused on using classical (e.g., [31,36]) and increasingly deep learning classifiers [39] to pick up these differences and determine age or other biometric traits from the EEG recordings. For example, to classify the age of participants based on the EEG features, researchers have used RF [36] and SVM-RBF [31], as well as DNNs [39]. The DNN method was a variant of long-short-term memory (LSTM) architecture and reached an age prediction accuracy of 93.7%. In the BCI applications, however, the goal is to reduce the performance dependency on biometric traits. Many studies have already looked at whether and how aging affects BCI performance. For example, it has been shown that the accuracy of the LDA classifier in vibro-tactile EEG BCI tasks was more than 15.9% lower for older subjects, as compared to that of the younger subjects [30]. Similarly, elderly people had slightly worse BCI performance on an eight-target EEG BCI spelling interface using code-modulated visual evoked potentials as features [44]. Another study tested the age-related drop in performance on the EEG BCIs with visual stimulations and also showed lower performance in the elderly when they used motion-onset VEP and steady-state-response VEP (SSMVEP & SSVEP) as features [45]. Lastly, Volosyak et al. [46] demonstrated a significant difference between the performance of younger and older subjects, again using the SSVEP features. The accuracy was 98.49% for young and 91.13% for older participants. While the presented studies clearly demonstrate the age-related differences, they only vaguely point to the possible reasons and solutions. Our goal was to dig deeper into the analysis of the classification results to obtain an explanation of what is a direct consequence of observed performance differences. Rather than focusing on an increase in predictive power, we tried to focus on the explainability of the results. With that in mind, we also utilized a great time resolution of the EEG and pointed to the possible caveats and solutions when the same classifier is used on people with different biometric traits, such as age.

2. Methods

The general workflow of the study is graphically depicted as a diagram in Figure 1. In this section, the general steps are further explained in detail.

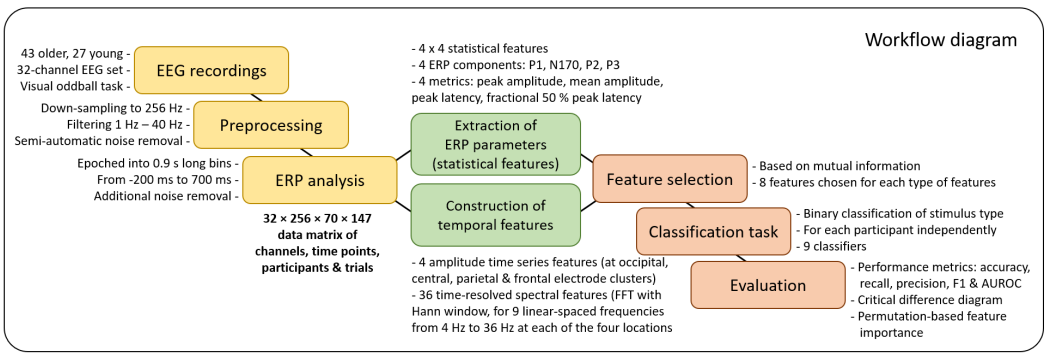

Figure 1. Workflow diagram.

2.1. Participants

Data (n = 79) for this analysis were collected in the framework of two different studies. Data from the first study included 46 older participants, 3 of whom were excluded during preprocessing stage due to excessive non-brain-related artifacts. Data from the second study comprised 33 young participants, 2 of whom were initially excluded due to nausea during EEG measurements, and 4 later due to non-brain-related artifacts. Thus, the final classification task included data from 43 older participants (27 females, mean age = 67.4 ± 5.5 years) and 27 young participants (14 females, mean age = 34.2 ± 2.3 years). The average number of years of education was 13 for the older participants and 16 for the younger participants. None had a history of psychiatric or neurological disorders, and all reported normal or corrected-to-normal vision. To exclude for mild cognitive impairment (MCI), the Montreal Cognitive Assessment (MoCA) screening tool was administered to the older group. All older participants scored above the threshold for MCI (>26/30 points), as was defined by Nasreddine et al. in their original paper [47]. All procedures were carried out in accordance with the ethical standards of the 1964 Declaration of Helsinki and were approved by the National Medical Ethics Committee (No. KME 57/06/17). Written informed consent was obtained from all participants prior to study enrollment.

2.2. Visual Oddball Task

A simple visual two-stimuli oddball task (see Figure 2) was used to obtain EEG activity of visual stimulus processing. Participants were seated in front of a 17-inch computer screen, at a 50 cm distance. On a black background, a white empty square (4.1 cm^2) was displayed in the center. Inside the square, 150 ms long stimuli appeared throughout the task. The task involved the presentation of 124 (84%) frequent non-target stimuli (white solid square, 4.1 cm^2) and 23 (16%) rare, randomly occurring, target stimuli (Einstein's face, 4.1 cm^2), with on average 669.6 ms interstimulus interval (SD = 11.8 ms). Participants were instructed to silently count the number of times the target stimulus occurred and report the sum at the end of the presentation. No motor response was required. While it has been shown that the ERP waveform to the visual oddball paradigm differs between overt (keypress) and covert (silent counting) conditions, the components differed only in their magnitude. Based on the topography and dipole modeling, the same brain areas become activated [48]. While we saved the behavioral results, which was the sum of the target trials at the end of the task, we have not used it in the analysis, as all participants correctly counted the number of target stimuli, except for one participant, who missed the correct sum by one.

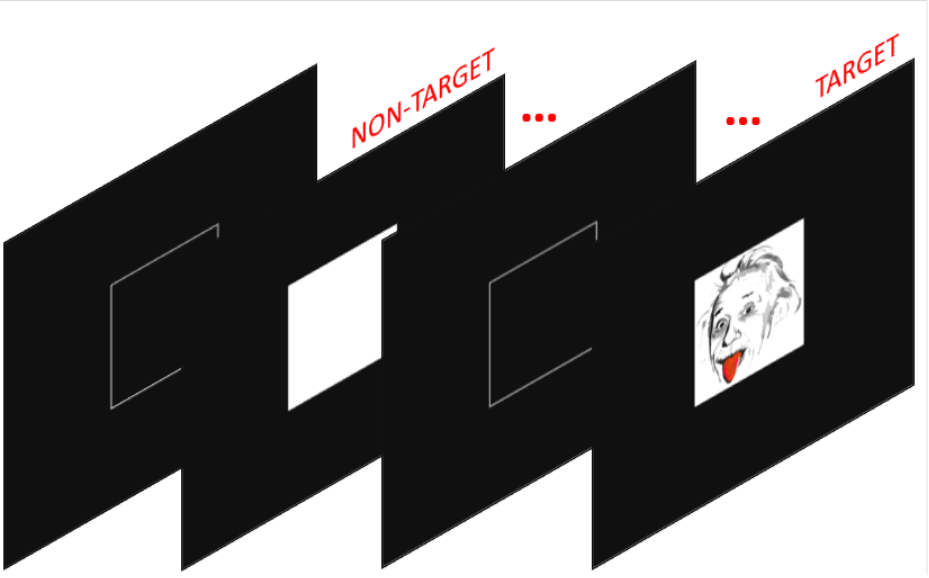

Figure 2. Visual oddball task.

2.3. EEG Analysis

EEG activity in the visual oddball task was recorded using a 32-channel EEG set consisting of 32 Ag/AgCl electrodes arranged according to the international 10–20 system. The signal was recorded at a frequency of 512 Hz and a resolution of 32 bits. The collected data were preprocessed using EEGLAB, an open source signal processing software [49] within the MATLAB program [50]. Recordings were visually inspected and high-noise sections were manually removed. Data were down-sampled to 256 Hz, filtered using 1 Hz high-pass and 40 Hz low-pass in-built EEGLAB filters and re-referenced to average. Using the Clean Rawdata plug-in, each channel was seemingly interpolated and correlated to its neighboring channels. All channels which had a correlation to their neighboring channels below 0.85 for more than half of the recording time, were then actually interpolated. Independent component (IC) analysis was used to obtain eye and muscle movement-specific information, as well as non-brain related information (e.g., line noise), which was then removed with the help of ICLabel classifier [51]. All ICs that were classified as the non-brain-related category above 85% certainty were removed from further analysis.

2.4. ERP Analysis

The continuous preprocessed data were then epoched into 0.9 s long bins that started 200 ms before the stimulus. Additionally, we removed epochs with an absolute peak amplitude over 100 µV. If more than 25% of all epochs were removed, the entire recording was excluded from the study. The minimum number of retained trials per participant was 129 (87.8%) trials, while the average was 145.9 trials (99.2%).

2.5. Extraction of Temporal Features

Forty features were calculated from the multi-dimensional $32 \times 256 \times 70 \times 147$ (channels, time points, participants, trials) data matrix. Four features corresponded to the voltage levels of four central spatial clusters at the channel level. The occipital cluster represented the average voltage fluctuations at O1, Oz and O2 channels, the parietal at P3, Pz and P4 channels, the central cluster C3, Cz and C4 channels and the frontal represented the average of F3, Fz and F4 channels. The remaining 36 features corresponded to the estimated spectral power (in dB) at 9 linear-spaced frequencies from 4 Hz to 36 Hz, measured at all four electrode clusters. The power was calculated based on the time-frequency analysis using the Fast Fourier spectral decomposition with a sliding Hann window of 32 time points. Analysis was performed using EEGLAB. To obtain power estimates of sufficient quality, 16 time points were removed from each end of the time series. The new reduced time interval between −136 ms and 734 ms (223 time points) was consequently used. Further mathematical details of the temporal feature construction can be seen in Appendix B. Additionally, we calculated mutual information between the features and a target and used it as a feature selection method to decrease the number of features to the best eight, as shown in Appendix D. The filter selection method was chosen because it is model agnostic and less prone to overfit on small datasets. In summary, the data in the classification task were composed of 10,214 trials (split between 70 participants) $\times$ 223 time points $\times$ 8 features. Selected features in the final dataset were occipital and central amplitudes, power at 4 Hz and 8 Hz in the occipital channels, power at 4 Hz, 8 Hz, and 12 Hz in the central channels, and power at 4 Hz in the frontal channel cluster. Figure 3 shows the trial-averaged features, grouped by age and stimulus type.

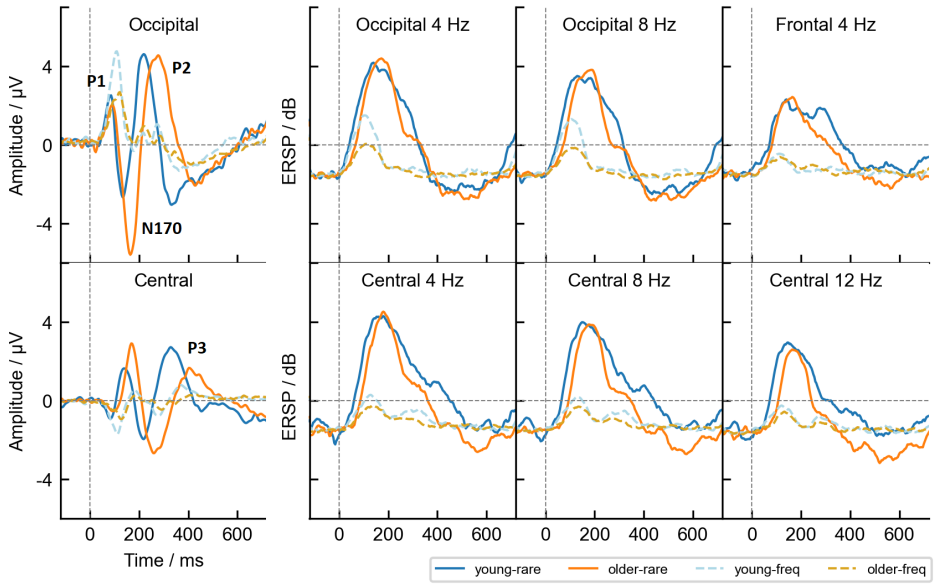

Figure 3. Trial-averaged temporal features grouped by age and stimulus type. The groups are distinguished by color as depicted in the legend. The first column displays amplitudes for the occipital electrode cluster (top row) and the central electrode cluster (bottom). In these subplots, the ERP components used in the second dataset are annotated in bold. The remaining subplots show changes in spectral power over time, at frequencies specified in the title of each plot. A vertical gray dashed line at 0 ms marks the onset of stimulus presentation. ERSP: Event-related spectral perturbation, shown in decibels (dB).

2.6. Extraction of Time-Independent Statistical ERP Features

The same multi-dimensional 32 × 256 × 70 × 147 (channels, time points, participants, trials) data matrix was used to create the second dataset. For every subject, the most relevant four visual ERP components were identified: P1, N170 and P2 from the occipital electrode cluster and the P3 component from the cluster of central electrodes, as marked in Figure 3. Each component was parameterized with four measures: peak amplitude, mean amplitude, peak latency and fractional 50% peak latency, all within a certain time interval. The P1 amplitudes and latencies were extracted within 50–150 ms interval, the N170 within 100–200 ms interval, the P2 within 200–325 ms interval and the P3 within 250–500 ms interval. Time intervals were chosen based on the previous research and the collapsed localizers approach [52], that is the zero-crossings of age group and stimulus averaged ERP waveform. Further mathematical details of the time-independent feature construction can be seen in Appendix B. As in the first dataset, the number of features was reduced from the original 16 features to 8 features, as shown in Appendix D. Based on the mutual information, both latency measures were found to be superior to amplitude measures. Nevertheless, instead of just highly correlated latency measures (see Appendix D for visual comparison), we decided to keep the best amplitude and the best latency measure, which were consistently the peak amplitude and the fractional 50% peak latency. All in all, the second dataset was comprised of 10214 trials (split between 70 participants) × 8 features. The features were the peak amplitudes and the fractional 50% peak latencies at 4 electrode clusters (occipital, parietal, central and frontal). Figure 4 shows the chosen trial-averaged features, grouped by age and stimulus type. The mean amplitudes (MA) and the peak latencies (PL), which were not included in classification tasks, are shown in Appendix A.

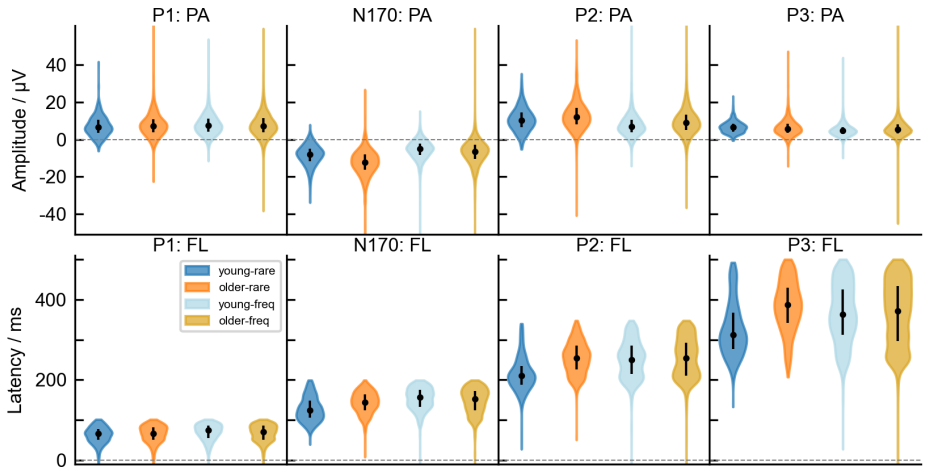

Figure 4. Trial-averaged time-independent features grouped by age and stimulus type. The groups are color-coded as indicated in the legend. The peak amplitudes (PA) are shown in the top row and the fractional 50% peak latencies (FL) are shown in the bottom row. The P1, N170, P2 components were calculated from the occipital electrode cluster, while the P3 component from the central electrode cluster.

2.7. Classification Task

The task was a single-target classification with the stimulus type (frequent/rare) as the binary target. Among the trials, 124 (84%) trials were labeled as frequent (presentation of a white square) and 23 (16%) as rare (presentation of Einstein's face) per participant. In the case of temporal features, classification was performed on each participant and time point independently, while in the case of time-independent statistical ERP features, time was not considered. The data were first split using a 10-fold cross-validation procedure. During each of the ten folds, the training data were then balanced for stimulus type to avoid majority class bias. The largest class (trials with frequent stimuli) was randomly undersampled to the average of both classes, while the minority class (trials with rare stimuli) was oversampled to the same value using SMOTE, Synthetic Minority Oversampling Technique [53]. That procedure allowed training the classifier on two equally sized classes, each with 66 trials. The training data were then shuffled. Finally, before training, the min-max normalization was performed on the training folds and applied on both train and test sets.

Six classical machine learning algorithms were tested on the task—three linear, namely LDA, SVC with linear kernel (SVC-Lin), and LR, three nonlinear, namely SVC with RBF kernel (SVB-RBF), KNN (k = 3) and a tree, as well as three ensemble methods: RF, AdaBoost and XGB, each set to 100 estimators and the maximum depth of 4. If not otherwise noted, the classifiers were applied with default parameter settings, either from the scikit-learn [54] or the xgboost module [55] in the case of XGB. Deep learning methods were not evaluated on this task due to the typically large number of parameters that would need to be optimized, for which insufficient training data were available.

Classifiers were evaluated based on their accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic (AUROC) metrics. Further mathematical details of how the evaluation metrics were calculated can be seen in Appendix C. Although we show the results for all metrics, we focus primarily on the AUROC score due to the highly unbalanced dataset [56,57]. The classifiers were statistically evaluated based on the AUROC scores of each participant using the critical difference diagram [58]. Critical difference measure includes the Friedman test with the corresponding post hoc Wilcoxon tests for pair-wise comparison between the classifiers. Statistical results were corrected for multiple comparisons using Holm's method [59]. Age-related statistical comparisons were carried out using ANOVA and post hoc independent t-tests. Lastly, feature importance was

calculated using a permutation-based technique to assess the relative contribution of each feature to the classifiers' performance.

3. Results

3.1. Differences between Classifiers

Figure 5 shows the performance scores (accuracy, AUROC, precision, recall, and F1) for each dataset type and classifier separately, averaged among participants. The results for the dataset with time-independent features are depicted by the filled circles around 0 ms, while the results for the dataset with temporal features span over the time axis as a line plot. Firstly, as expected, we observed a noticeable discrepancy between accuracy and AUROC metrics. The accuracy scores were much higher compared to AUROC scores. For example, the best accuracy scores for RF were 86.7% with the time-independent features and 78.4% with the temporal features, while the AUROC scores were 78.9% and 71.3%, respectively. Another discrepancy was that almost all classifiers had above-chance accuracy even before the stimulus appeared (e.g., RF had an average accuracy rating of around 65% before 0 ms). As already mentioned in the Methods, the most likely reason for such a non-intuitive result is the influence of imbalanced datasets. The accuracy score is highly affected by the imbalance between the classes, as the score of the more prominent class overshadows the score of the less represented class, in which we are usually more interested in [60]. Although undersampling and oversampling techniques mitigated class inequalities, the test set was not balanced, which also greatly affected the performance evaluations.

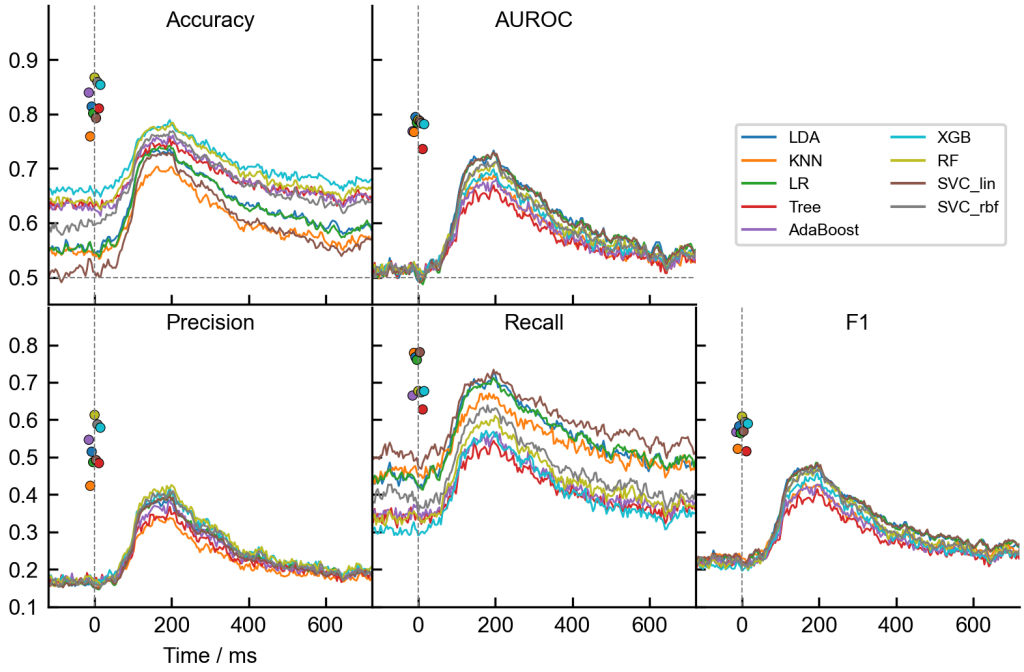

Figure 5. Classifiers' performances based on five metrics: accuracy, AUROC, precision, recall and F1 score. Results for both datasets are presented. The lines show the results for the dataset with temporal features, averaged over participants. The filled circles represent the averaged results based on the time-independent features. These results, also averaged over participants, do not include time dimension and are jittered around 0 ms just for presentation purposes. Classifiers are color coded as shown in the legend. The horizontal gray dashed line at 0.5 indicates the threshold level at which the binary classifier performs at random. The vertical gray dashed line at 0 ms marks the onset of the stimulus presentation.

This interpretation is also supported by the precision and recall scores, or their F1 weighted average metric. All the classifiers had a poor precision score, meaning that there

were relatively many false positives (frequent stimuli identified as rare). As an example, the RF had a precision score of 61.3% with the time-independent features and 42.6% with the temporal features. On the other hand, the recall scores were much better but also varied significantly between the classification models. Recall defines the ability of the classifiers to detect rare stimuli without many false negatives (rare stimuli identified as frequent). Maximum recall scores for RF were 67.7% with the time-independent features and 61.1% with the temporal features. In sum, recall/precision scores show that rare stimuli were rarely misclassified, while frequent stimuli were misclassified more often. We believe the misclassification of more numerous frequent trials gave false positives more weight in the accuracy score calculation and consequently reduced precision. The reason for the high number of false positives could also be greater noisy voltage fluctuations of single trials. As the AUROC score alleviates some of the drawbacks of the accuracy measure and is generally a better metric due to its higher discriminating power [56,57], it is the primary metric we focused on in this paper.

Secondly, all classifiers generally showed relatively low performance and moreover, classifiers showed much lower performance when trained on temporal features as compared to the training on time-independent statistical features. For example, Figure 5 shows that the best-performing classifiers were linear classifiers LDA, LR and SVC-Lin. The maximum LDA AUROC score on the time-independent dataset was 79.5%, while it was 73.3% on the temporal dataset. Similarly, the worst-performing single decision tree had a score of 73.6% for the time-independent dataset, while only 66.9% for the temporal dataset. The most likely reason for such difference is that the extraction of the relevant characteristics from the time series retains most of the information needed for the classifiers and additionally improves the robustness by removing the time variability in the data.

A third observation from Figure 5 is the temporal evolution of the decodable signal. The line plots show how approximately 50 ms after the stimulus presentation the classifiers were able to detect stimulus-related differences in the EEG signal above chance. This indicated that there were already relevant P1 component differences, enabling classifiers to separate the trials. However, the best time interval to infer the stimulus type was around 100–250 ms, which mainly corresponds to the P1 and N170 visual evoked potentials, based on the voltage fluctuations at the sensor level (see Figure 3). The performance peaks of the best three linear classifiers (LDA, LR and SVC-Lin) occurred at approximately 195 ms. After that, the performance of all models drops gradually until the end of the trial.

To statistically assess the performance of the classifiers, we performed critical difference diagram calculations using the Friedman test with the post-hoc analysis. In Figure 6A, the critical difference diagram is shown for the dataset with the time-independent ERP features. The critical distance value was calculated to be between 1.34 and 1.56. The single worst-performing classifier was a decision tree. Besides that, the three nonexclusive clusters emerged. The best-performing cluster included mainly linear classifiers, with LDA being ranked first. In Figure 6B, we extended the critical difference diagram to two dimensions, to show how the differences between classifiers vary over time. The colored markers on top of the gray background show how one of the four representative classifiers, LDA, RF, XGB or Tree, compared pairwise to all the rest. *No* marker at a specific time point represents *no* statistically significant difference ($p > 0.05$) on a multi-group Friedman test, while a marker at a specific time point represents *no* significant difference on the post-hoc pairwise test between the representative classifier and another model. In other words, statistically similar classifiers are grouped together by color. Although only four representative clusters are shown, they encompass most of the observed variability. The 2D results support the differences seen among the classifiers in Figure 5. It can be clearly observed that the classifiers did not significantly differ in their scores up to around 50 ms after the stimulus presentation. However, at approximately 100 ms, the four clusters emerged. The best cluster is represented by the LDA (marked in blue), which performed similarly well to LR and SVC-Lin. At the same time, these three models performed significantly better as all the rest, at least for several time points. Furthermore, we can observe that during the

same time period, KNN classifier performed significantly worse than RF (purple), but not significantly worse than the decision tree (brown). Results nicely align with subplot A and suggest that more robust, high-bias, linear classifiers performed significantly better at this task, especially within the 200–500 ms time window.

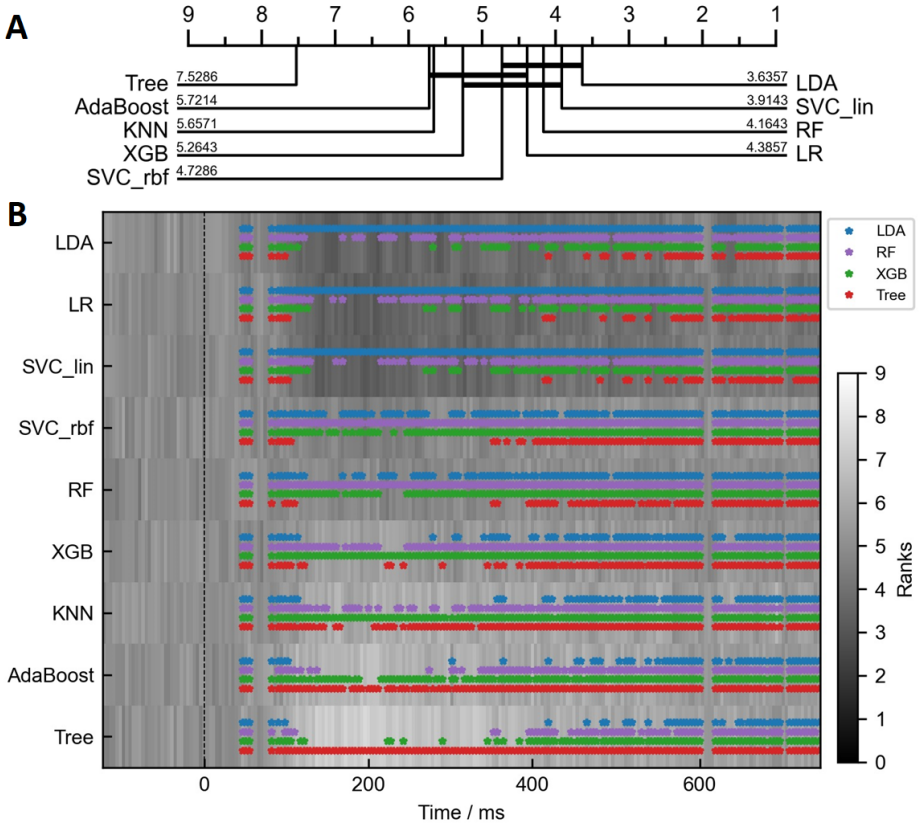

Figure 6. Statistical comparison of AUROC scores between classifiers using critical difference diagram. (**A**) The critical difference diagram compares the classifiers trained on time-independent features. (**B**) The comparison of the classifiers that were trained on each time point independently. The background color represents the rank of a specific classifier at a specific time point. The lower the rank (the darker the color), the better the classifier performed compared to others. If the markers on top are present, the Friedman test was statistically significant for a specific time point. The markers on top represent pairwise post hoc statistically *non-significant* differences ($p > 0.05$), meaning that at each time point, the similarly performing classifiers are grouped together. In this plot, only four representative clusters are shown, LDA (blue), RF (purple), XGB (green) and KNN (brown).

Finally, the duration of the classification tasks is shown in Table 1 for each of the datasets. The training using the larger dataset with temporal features took from 28 min for the fastest decision tree, LDA, and LR to the slowest ensemble classifiers, which took up to 325 min (RF). As expected, the training times were much faster for the time-independent dataset, for which the loop over the time points was not needed. Again, the fastest were the linear classifiers and the decision tree with less than half a minute, while the ensemble classifiers took the longest, up to almost 18 min. These results clearly show that the dataset with time-independent statistical features is practically the only viable option for real-time EEG BCI applications.

Table 1. Duration of classification tasks, measured in minutes.

	LDA	LR	SVC_lin	SVC_RBF	RF	XGB	KNN	AdaBoost	Tree
Temporal	29	29	47	63	325	110	139	313	28
Statistical	0.3	0.3	0.4	0.6	17.6	1.4	2	3.1	0.3

3.2. Age-Related Differences

Our main question addressed in this paper was related to the potential aging effects in stimulus type classification. First, we grouped the AUROC scores based on age, as shown in Figure 7 for three representative classifiers: LDA, RF and KNN. Representative classifiers were chosen based on their fundamental model structure, where LDA represented the linear, RF the ensemble and KNN the non-linear best-performing classifiers. It can be appreciated from the figure that the performance of the linear classifiers was not affected by age, while on the other hand, the ensemble and the non-linear classifiers showed decreased performance in the older age group. When using the dataset with temporal features, the differences occurred only during a short time interval between approximately 55–129 ms for RF and between 98–133 ms for KNN. However, when the compressed time-independent features were used, age highly influenced the classification performance overall. In order to make a better comparison between dataset types, we extracted participants' best AUROC scores based on temporal features and compared them with the above-mentioned scores from the time-independent dataset. The averaged maximum scores are quantitatively shown in Table 2 and visually depicted in Figure 8. After ANOVA tests confirmed group-level statistical differences ($p < 0.01$) for all classifiers, pairwise independent t-tests were performed. Again but now quantitatively, we confirmed significant age-related differences in classification performance for the dataset with time-independent features. The performance score was on average 0.828 (SD = 0.11) in the younger group and 0.765 (SD = 0.092) in the older age group when RF was used (marked by letter *c* in Table 2). Similarly, the score was 0.805 (SD = 0.118) for the young group, and 0.743 (SD = 0.098) in the older age group when KNN was used (marked by letter *e* in Table 2). Surprisingly, however, the maximum scores for the dataset with temporal features did not differ between age groups for any of the classifiers. Furthermore, we have additionally observed that all classifiers performed significantly better for older participants when the dataset with temporal features was used (see the letters a, b and d in Table 2. These results are to some extent contradictory to the observations from Figure 5, where it was clearly shown that the classification performance using the dataset with extracted statistical ERP features outperformed the performance using the dataset with temporal features. However, the important distinction is that in Figure 5, the individual maximum values were spread over several time points and averaging seemingly reduced the performance of such an approach. Another point to notice is the higher variance of the AUROC scores within the time-invariant dataset, as compared to the alternative, which can also be appreciated from the standard deviations in Table 2. On average, the standard deviations were 0.076 and 0.106 for temporal and time-independent features, respectively.

Table 2. Averaged individual participant's maximum AUROC scores grouped by age and dataset type. Standard deviation is shown in parentheses. Statistically significant pairwise differences ($p < 0.05$) are marked by bolded superscript letters.

		LDA	RF	KNN
Temporal	Young	0.860 (0.068)	0.848 (0.084)	0.846 (0.083)
	Older	**0.839** ^{a} (0.072)	**0.817** ^{b} (0.074)	**0.812** ^{d} (0.077)
Statistical	Young	0.808 (0.118)	**0.828** ^{c} (0.110)	**0.805** ^{e} (0.118)
	Older	**0.786** ^{a} (0.101)	**0.765** ^{b,c} (0.092)	**0.743** ^{d,e} (0.098)

Up to now, we have explored the age-related differences based on the AUROC metric. While the performance metric alone could not provide an answer as to why the linear models were not affected by age, permutation-based feature ranking results shown in Figure 9 provided some clues. When the LDA was trained on the temporal features, it used a variety of features in its classification task and more than the other two classifiers based its linear boundary on spectral features. The RF and KNN, on the other hand, mainly relied on the occipital and central cluster amplitudes. As the amplitudes notably differed between the age groups, their usage inevitably led to performance discrepancies. For example, the importance of the occipital amplitude for the RF was the strongest during the P1 and the N170 component (50–150 ms), as well as during the P2 component (at 235 ms) for the younger group. In the older group, however, the occipital amplitude was important only during one stronger peak at around 300 ms, representing the averaged P2 peak (see Figure 3). Similarly, we observed an age-related time shift in the importance of the central amplitude, where the timings nicely corresponded to the P3 peak in each age group—the central cluster was most important at 340 ms in the young and only at 400 ms in the older age group. Furthermore, when we examined the feature importance of the time-independent features, similar distinctions were found. The most important features in RF and KNN were the N170 fractional 50% peak latency in the young and the N170 peak amplitude together with P3 fractional 50% peak latency in the older age group. All of the features are good in discriminating the stimulus type but do carry differences between the age groups as well (see Figure 4). In the LDA model, however, besides the N170 features, the P2 peak amplitude, agnostic to age differences, was ranked as one of the most important features in both age groups.

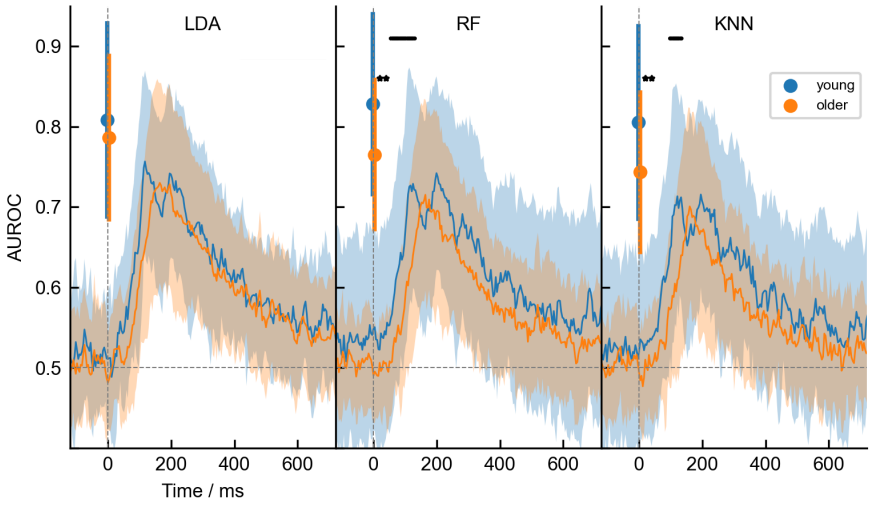

Figure 7. Age-related differences in AUROC scores of the three representative classifiers (LDA, RF, and KNN). The plots contain results for both datasets. The results for the training on the dataset with time-independent features are represented at 0 ms by filled circles representing the average and the error bars representing the standard deviation. Significant differences in the AUROC score between the young (blue) and older (orange) populations are shown by the two black stars ($p < 0.01$) next to the error bars. The results for the dataset with temporal features are presented as line plots. The thick line represents the average, while the shadowed area around represents the standard deviation over the trial duration. Black horizontal thick lines on the top mark the time intervals in which significant age-related differences in performance were observed. The vertical gray dashed line at 0 ms marks the onset of stimulus presentation, while the horizontal gray dashed line at 0.5 represents the chance level performance.

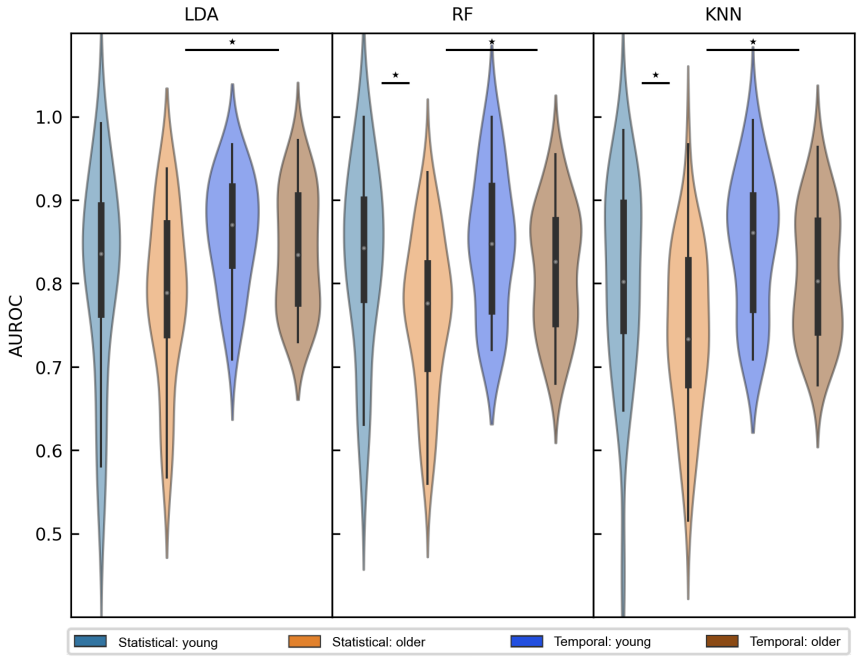

Figure 8. Individual subject's best AUROC scores, grouped by age (young/older) and classification dataset type, that either includes features across time (labeled *Temporal*) or time-independent ERP parameters (labeled *Statistical*). Results are shown for the three representative classifiers (LDA: **left**, RF: middle, KNN: **right** column). Significant difference between two groups is represented by a star ($p < 0.05$).

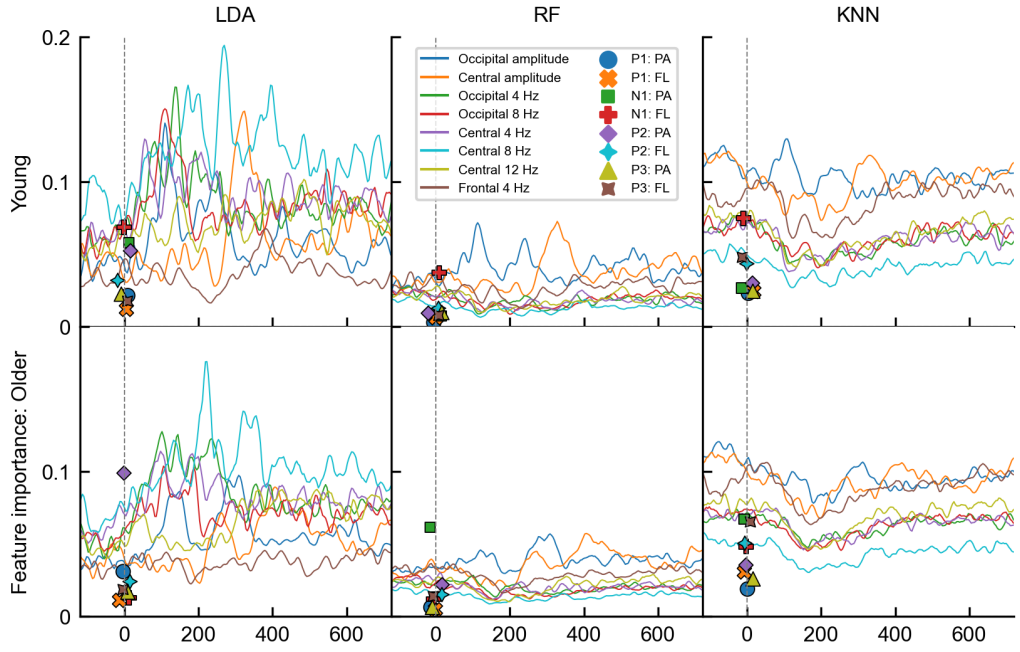

Figure 9. Permutation-based feature importance results. Each column shows the results of a representative classifier (LDA: **left**, RF: **middle**, KNN: **right** column). The line plots show temporal evolution of feature importances, separated for young (**top row**) and older participants (**bottom row**). The vertical gray dashed line at 0 ms marks the onset of stimulus presentation. The importances of time-independent features are depicted with various marker styles at 0 ms. Each feature is color coded as shown in the legend. Note that the N170 component is labeled as N1.

4. Discussion

In this study, we investigated the possible effects of aging on classification performance in discriminating rare and frequent trials in a passive visual oddball paradigm. To assess the aging effects more broadly, we used nine different classification models, as well as two types of datasets, one with statistical time-independent ERP features and another with temporal amplitude and spectral features.

Since many classification models were used in the analysis, we first compared their performances based on the AUROC score, which was most robust to imbalanced data. Generally, all commonly used machine learning algorithms have already been used in EEG classification tasks. A recent review has shown that, besides the deep learning algorithms, the support vector classifier (SVC) is the preferred supervised algorithm, performing with the highest accuracy [57,61]. LDA, KNN, and RF models were also noted as one of the best choices for the EEG signals. In our task, linear classifiers outperformed other classifiers, with LDA being ranked best. We assume this is because they form simpler linear decision boundaries that need fewer parameters to train, while more complex models, such as ensemble classifiers, overfit more easily when there are only a few training data available [57]. Such a result was expected, as the task included only a scarce number of especially target trials per participant. Additionally, linear models have high bias and thus the advantage when dealing with noisy and weak information signals, such as single-trial EEG data.

Next, we compared the performance between time-independent features and temporal features. In general, in BCI applications, processing speed is very important, so it is more efficient to use compressed time-independent features. On the other hand, temporal features can be used to study the temporal evolution of the signal and better understand feature importance [6]. Based on the results, the performance with the time-independent features seemed to be better at first glance, as we compared the average scores over the participants. This was expected since the extraction of relevant signal characteristics should preserve most of the information needed for the classifiers and additionally remove the time variability of the data. Interestingly, however, further analysis showed that the maximum AUROC scores of individual participants were comparable between datasets and were even significantly higher when the temporal features were used in the older age group. Another downside we observed with the time-independent features was a higher variance in performance. All in all, while the speed guarantees the usage of compressed time-independent features in the BCI applications, the presented results show that such feature extraction also brings some issues.

The primary issue we encountered was the choice of the time window for calculating the ERP parameters of each component, as the time interval can greatly affect the feature extraction and consequently the final results. One reason is that each of the ERP parameters prefers specific settings for optimal extraction [52]. The peak amplitude requires stretched measurement window to encompass inter-subject variability, while the mean amplitude functions better at shorter, specific time windows that only encompass the measured peak. This discrepancy can be also observed in our data. Although we followed well-accepted guidelines [52], high variability within as well as between the age groups led to very broad time windows. Consequently, the calculated mean amplitudes were all close to 0 µV (see Appendix A, even though the ERPs were prominent. Thus, in our case, the peak amplitude measure outperformed the mean amplitude, even though the mean amplitude is a rather preferred measure compared to the less noise-resistant peak amplitude. This has been also confirmed by the feature selection algorithm, which ranked the peak amplitude above the mean amplitude for all four components.

Another point to note is that the P3 was traditionally thought to be the most responsible component for the discrimination of target and non-target trials in the oddball study. While some studies still have such observations [57], others are showing that earlier VEP can also have the most discriminating power [62]. In our experiment, we observed that the trials can be discriminated above chance level as early as 50 ms after the stimulus presentation

and are best classified at around 195 ms, at the time of N170, which is much earlier than the emergence of the P3 component. We believe that the reason for such mixed results lies in the type of visual oddball task at hand. When a very different visual stimulus is presented for the target than it is for the non-target stimulus, there is a high probability that the difference will already be observed in the visual evoked potentials.

Moreover, it has been shown that the famous phenomenon of non-learners might be due to external reasons and that all people can learn to use at least one form of BCI [63]. The same authors also show that the demographic parameters, handedness, vision correction and BCI experience have no significant effect on the performance of VEP-based BCIs. What about aging?

Finally, we focused on the effect of aging on classification performance. While the suggested reasons for the poorer performance of the older age group have been the longer reaction times and slower learning [46], we show that aging per se does not necessarily affect classification performance. Instead, the effects of aging depend on the choice of the classifier and, to some extent related, on the choice of features. If the most informative features for the classifier contain age-related differences, this will most likely lead to classification differences as well. In our dataset, ERP waveforms show the clear age-related amplitude and latency differences already at the sensor level. Some classifiers, such as LDA, used features, which were less affected by age and consequently the performance between the age groups did not differ. Moreover, LDA picked up important information from various features and thus made more robust estimates by reducing potential within-class differences, such as age. To conclude, by the choice of features that are effectively discriminating stimulus type (or any other class of interest) but are agnostic to the possible within-class differences such as aging, we can increase the classification reliability for all participants. This remark is progressively more important, as we more commonly merge data of all participants to obtain larger training sets for larger and more complex classifiers. Consequently, the training is not participant-specific, but rather cross-participant transfer learning is needed and assumed. In such cases, the feature importance of the used classifier should be checked to avoid potential performance loss in a certain subclass due to age-related or other subclass-specific characteristics.

There are also weaknesses of the study that needed to be taken into account. Besides the standard classifiers, EEG-specific classifiers have emerged in recent years (e.g., adaptive classifiers) to account for the noisiness and non-stationarity in the signal, as well as limited training data [6]. While the adjusted models would most likely increase the single-trial classification performance, we believe that the effects of aging, which were investigated here, would not differ substantially. Additionally, the data were preprocessed using the computationally costly ICA, discarding the signal related to eye and muscle artifacts. Again, however, as mostly the posterior middle electrodes were selected as the best features, we believe the eyes and muscle-related artifact would not induce larger changes in the final conclusions. The greater limitation of the experiment and analysis is that participants completed relatively few trials, which reduced the signal-to-noise ratio and decreased the learning capability of the classifiers, especially the more complex ones that would need more training data. The aging effects are planned to be further examined on a larger dataset, which would also enable us to examine the potential effects on the source-based features.

Author Contributions: Conceptualization, N.O. and U.M.; Data curation, U.M.; Formal analysis, N.O.; Funding acquisition, U.M.; Investigation, M.P. and N.O.; Methodology, N.O. and U.M.; Project administration, U.M.; Writing—original draft, N.O.; Writing—review & editing, M.P., A.M., V.K., S.D. and U.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Slovenian Research Agency (research core Fundings No. P5-0381 and No. P2-0103) and European Social Fund and the Ministry of Education, Science and Sport (Slovenia, EU). The authors also acknowledge financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952401 (TwinBrain—TWINning the BRAIN with machine learning for neuro-muscular efficiency).

Institutional Review Board Statement: The procedures were carried out in accordance with the ethical standards of the 1964 Declaration of Helsinki and approved by the Slovenian National Medical Ethics Committee.

Informed Consent Statement: Written informed consent was obtained from all participants prior to inclusion in the study.

Data Availability Statement: Raw and preprocessed data, as well as datasets directly used in the classification tasks are located https://doi.org/10.5281/zenodo.7495536 (accessed on 30 January 2022). The code used for the analysis is available https://github.com/NinaOmejc/VEP_classification_aging.git (accessed on 30 January 2022).

Acknowledgments: The authors thank all participants for their time commitment and research assistants (David Medica, Dino Manzioni, Kaja Teraž, Daša Gorjan, Tina Skrt, Saša Bele, Maja Pušnik, Urška Jenko, and Tina Čeh) for their assistance with the study. The authors thank Boštjan Gec for the help with mathematical descriptions and prof. Gorazd Drevenšek (University of Primorska) for the use of EEG equipment and the staff from the Center for daily activities for the elderly Koper, Slovenia (EU) for all provided support and management.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUROC	Area under the receiver operating curve
BCI	Brain-computer interfaces
EEG	Electroencephalogram
ERP	Event-related potential
ERSP	Event-related spectral perturbation
FN	False negative
FP	False positive
FPR	False positive rate
FL	Fractional 50% peak latency
LDA	Linear discriminant analysis
LR	Linear regression
KNN	K-Nearest neighbors
MA	Mean amplitude
MCI	Mild cognitive impairment
PA	Peak amplitude
PL	Peak latency
RF	Random forest
SVC	Support vector machine
TN	True negative
TP	True positive
TPR	True positive rate
XGB	Extreme Gradient Boosting

Appendix A

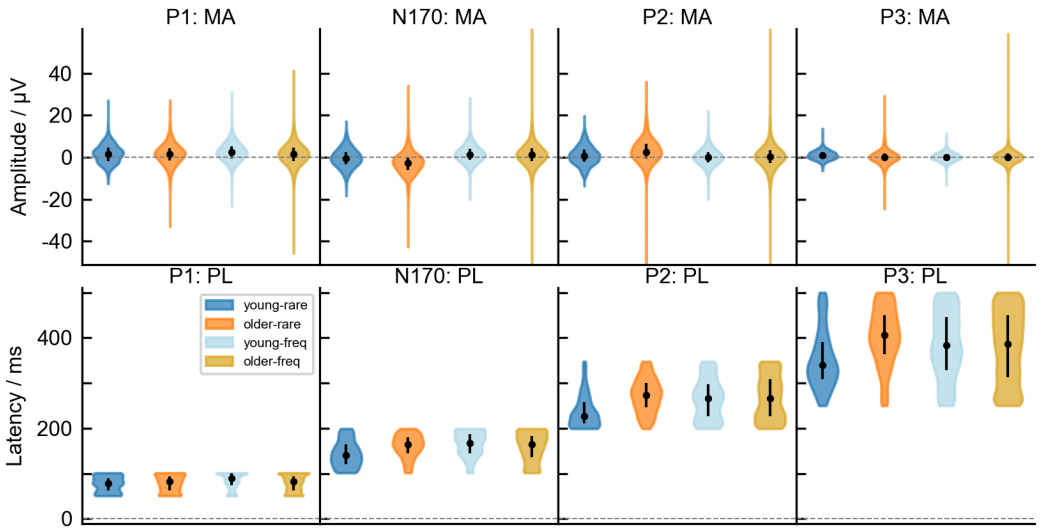

Figure A1. Trial-averaged time-independent features, which were not used in the classification task. Features are grouped by age and stimulus type, as color-coded by the legend. The mean amplitudes (MA) are shown in the top row and the peak latencies (PL) in the bottom row.

Appendix B

This Appendix describes the extraction of features in more detail. First, out of 32 electrodes, we calculated the averages of three electrodes at four brain regions, namely occipital (electrodes O1, Oz, O2), parietal (P3, Pz, P4), central (C3, Cz, C4), and frontal (F1, Fz, F2) region. The four averaged time series were then used in the next steps. First, the construction of the temporal features and then the construction of the statistical, time-independent features are presented.

Four of the 40 temporal features were directly the averaged amplitude signals at four locations as explained above, while the other 36 represented the event-related spectral perturbations (ERSP) for 9 discrete frequencies at each of the 4 locations. The procedure to calculate ERSP for the i-th sliding time window is shown in Equations (A1)–(A4).

$$w(n) = 0.5(1 - cos(2\pi\frac{n}{N})),\ 0 \leq n \leq N \tag{A1}$$

$$y_i(n) = w(n)\, x_i(n) \tag{A2}$$

$$Y_i(k) = (FFT(y_i))(k) \tag{A3}$$

$$ERSP_i(k) = |Y_i(k)|^2 \tag{A4}$$

where $w(n)$ generates the coefficients of the Hann window with the window length $N+1$ (Equation (A1)). In Equation (A2), the $y_i(n)$ represents the windowed signal, calculated as the multiplication between the Hann window and the $x_i(n)$, which represents the data in the time domain, or in other words, the voltage fluctuations of an electrode cluster at discrete time points n within the i-th time window. Then, Equation (A3) describes the power spectrum $Y_i(k)$ over the same window i, obtained by the Fast Fourier transform (FFT), where k is the frequency index. Finally, ERSP (in dB) is calculated as the power spectral density at frequency k for a single sliding window i (see Equation (A4)). Normally, as it is also in our case, the ERSP was calculated for all time points n, shifting the sliding window respectively.

Secondly, we calculated the ERP parameters that were used as time-independent features in the classification task. Calculations are described in Equations (A5)–(A8).

$$PL(t_{start}, t_{end}) = \begin{cases} argmax_{t(n)\in[t_{start},t_{end}]} x(n), & \text{if } ERP\ component\ is\ positive \\ argmin_{t(n)\in[t_{start},t_{end}]} x(n), & \text{if } ERP\ component\ is\ negative \end{cases} \quad (A5)$$

$$PA(t_{start}, t_{end}) = x(n),\ where\ t(n) = PL \quad (A6)$$

$$MA(t_{start}, t_{end}) = \frac{1}{n_{end} - n_{start}} \sum_{j=n_{start}}^{n_{end}} x(n_j),\ where\ t(n_{start}) = t_{start}\ \&\ t(n_{end}) = t_{end} \quad (A7)$$

$$FL(t_{start}, t_{end}) = t(n),\ where\ x(n) = \frac{1}{2} PA\ \&\ t(n) < PL \quad (A8)$$

Equation (A5) calculates the peak latency (PL), which is, depending on the ERP component polarity, a latency either at the voltage peak (for positive) or at the trough (for negative polarity) of a specific ERP component signal $x(n)$. P1, P2, and P3 components have positive polarity, while N170 has negative polarity. The function $t(n)$ represents the times of the signal at discrete time points, denoted by index n. The beginning (t_{start}) and the end (t_{end}) of the time interval in which the component is expected to be found, are empirically chosen for the individual components. Equation (A6) calculates the peak amplitude (PA) of a specific component (that is expected to occur between t_{start} and t_{end}). PA is defined as the amplitude of the signal $x(n)$ at the time of PL. Equation (A7) calculates the mean amplitude (MA), which is simply the mean of the signal within the time interval t_{start} and t_{end}. Lastly, Equation (A8) calculates the fractional 50% peak latency (FL). The metric finds the time $t(n)$ that occurs before the PL and at which the signal $x(n)$ reaches half of the PA.

Appendix C

We have used several evaluation metrics of classification performance, specifically accuracy, precision, recall, F1, and area-under-receiver-operating-curve (AUROC) scores. The following appendix section provides mathematical details and a short description of how the metrics were computed using the scikit-learn library.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (A9)$$

$$Precision = \frac{TP}{TP + FP} \quad (A10)$$

$$Recall = Sensitivity = TPR = \frac{TP}{TP + FN} \quad (A11)$$

$$Specificity = 1 - FPR = \frac{TN}{FP + TN} \quad (A12)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \quad (A13)$$

$$AUROC = \int_0^1 TPR\, d(FPR) \quad (A14)$$

Legend: FN: False negative, FP: False positive, FPR: False positive rate, TN: True negative, TP: True positive, TPR: True positive rate.

The Equation (A9) calculates the accuracy of classification, which is the ratio of correct classifications (true positives and true negatives) versus all classifications. Precision (A10) is the ratio of correctly identified (true) positive classifications (TP) versus all classifications, that were (correctly or falsely) classified as positive. Recall or the true positive rate (TPR) is defined in the Equation (A11) as the ratio of correctly classified positive classifications (TP)

and the sum of true positive and false negative classifications (which sums to all samples, which should have been classified as positive). Specificity (A12) is a ratio of correctly classified negative classifications (TN) and the sum of true negative and false positive classifications (which sums to all negative samples). F1 score (A13) is the harmonic mean of precision and recall, where both metrics have equal weight. Lastly, the AUROC score (A14) is defined as the integral of (or the area under) the ROC curve, which is a probability curve, that tells how much the classifier is capable of distinguishing between classes depending on the threshold between recall and specificity, by which we can adjust the number of type 1 (false-positive) and type 2 (false-negative) errors.

Appendix D

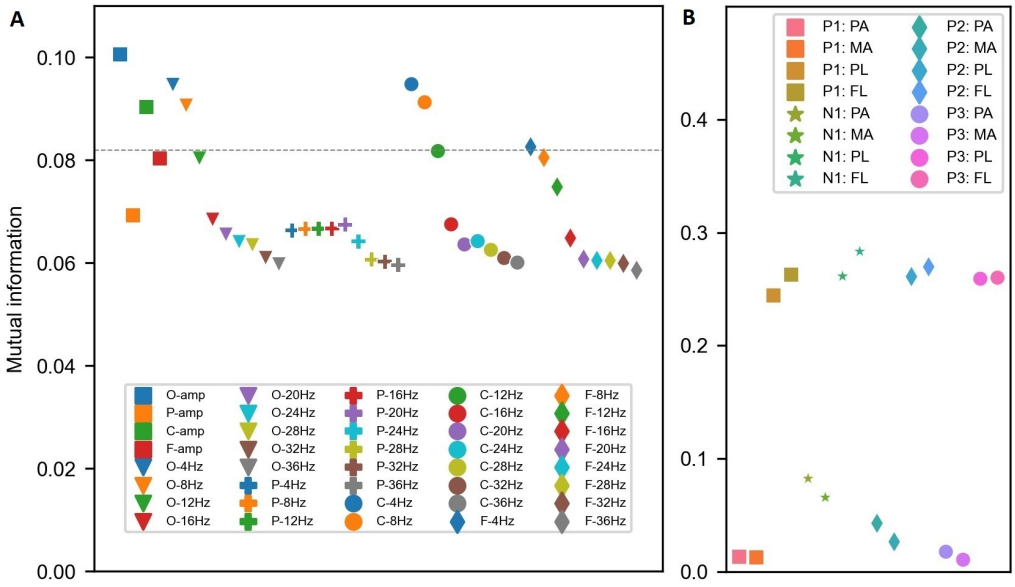

Figure A2. Feature selection based on mutual information. (**A**) Selected eight best temporal features are above the gray horizontal dashed line. (**B**) Selection of eight best time-independent statistical ERP features. The best latency and amplitude measures were selected, which were consistently the peak amplitude and the fractional 50% peak latency measures. Note that the N170 component is labeled as N1.

References

1. Rashid, M.; Sulaiman, N.; Majeed, A.P.P.A.; Musa, R.M.; Nasir, A.F.A.; Bari, B.S.; Khatun, S. Current status, challenges, and possible solutions of EEG-based brain-computer interface: A comprehensive review. *Front. Neurorobot.* **2020**, *14*, 25. [CrossRef]
2. Breakspear, M. Dynamic models of large-scale brain activity. *Nat. Neurosci.* **2017**, *20*, 340–352. [CrossRef]
3. Ghorbanian, P.; Ramakrishnan, S.; Ashrafiuon, H. Stochastic non-linear oscillator models of EEG: The alzheimer's disease case. *Front. Comput. Neurosci.* **2015**, *9*, 48. [CrossRef] [PubMed]
4. Miladinović, A.; Ajčević, M.; Jarmolowska, J.; Marusic, U.; Colussi, M.; Silveri, G.; Battaglini, P.P.; Accardo, A. Effect of power feature covariance shift on BCI spatial-filtering techniques: A comparative study. *Comput. Methods Prog. Biomed.* **2021**, *198*, 105808. [CrossRef]
5. Wolpaw, J.R.; Birbaumer, N.; McFarland, D.J.; Pfurtscheller, G.; Vaughan, T.M. Brain-computer interfaces for communication and control. *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.* **2002**, *113*, 767–91. [CrossRef]
6. Lotte, F.; Bougrain, L.; Cichocki, A.; Clerc, M.; Congedo, M.; Rakotomamonjy, A.; Yger, F. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *J. Neural Eng.* **2018**, *15*, 031005. [CrossRef]
7. Bamdad, M.; Zarshenas, H.; Auais, M.A. Application of BCI systems in neurorehabilitation: A scoping review. *Disabil. Rehabil. Assist. Technol.* **2015**, *10*, 355–364. [CrossRef]
8. Aggarwal, S.; Chugh, N. Review of machine learning techniques for EEG based brain computer interface. *Arch. Comput. Methods Eng.* **2022**, *29*, 3001–3020. [CrossRef]
9. Luck, S.J. *Event-Related Potentials*; American Psychological Association: Washington, DC, USA, 2012.
10. Daniel, S.; Bentin, S. Age-related changes in processing faces from detection to identification: ERP evidence. *Neurobiol. Aging* **2012**, *33*, 206.e1–206.e28. [CrossRef]

11. Luck, S.J. *An Introduction to the Event-Related Potential Technique*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2014; p. 416.
12. Dustman, R.E.; Beck, E.C. The effects of maturation and aging on the wave form of visually evoked potentials. *Electroencephalogr. Clin. Neurophysiol.* **1969**, *26*, 2–11. [CrossRef]
13. Polich, J. EEG and ERP assessment of normal aging. *Electroencephalogr. Clin. Neurophysiol. Potentials Sect.* **1997**, *104*, 244–256. [CrossRef] [PubMed]
14. Grady, C.L. Cognitive neuroscience of aging. *Ann. N. Y. Acad. Sci.* **2008**, *1124*, 127–144. [CrossRef] [PubMed]
15. Anderson, A.J.; Perone, S. Developmental change in the resting state electroencephalogram: Insights into cognition and the brain. *Brain Cogn.* **2018**, *126*, 40–52. [CrossRef] [PubMed]
16. Salthouse, T.A. The processing-speed theory of adult age differences in cognition. *Psychol. Rev.* **1996**, *103*, 403. [CrossRef]
17. Vysata, O.; Kukal, J.; Prochazka, A.; Pazdera, L.; Valis, M. Age-related changes in the energy and spectral composition of EEG. *Neurophysiology* **2012**, *44*, 63–67. [CrossRef]
18. Celesia, G.G.; Kaufman, D.; Cone, S. Effects of age and sex on pattern electroretinograms and visual evoked potentials. *Electroencephalogr. Clin. Neurophysiol. Potentials Sect.* **1987**, *68*, 161–171. [CrossRef]
19. Mitchell, K.; Howe, J.; Spencer, S. Visual evoked potentials in the older population: Age and gender effects. *Clin. Phys. Physiol. Meas.* **1987**, *8*, 317. [CrossRef]
20. Aleksić, P.; Raicević, R.; Stamenković, M.; Djordjević, D. Effect of aging on visual evoked potentials. *Vojnosanit. Pregl.* **2000**, *57*, 297–302.
21. Kuba, M.; Kremláček, J.; Langrová, J.; Kubová, Z.; Szanyi, J.; Vít, F. Aging effect in pattern, motion and cognitive visual evoked potentials. *Vis. Res.* **2012**, *62*, 9–16. [CrossRef]
22. Kropotov, J.; Ponomarev, V.; Tereshchenko, E.P.; Müller, A.; Jäncke, L. Effect of aging on ERP components of cognitive control. *Front. Aging Neurosci.* **2016**, *8*, 69. [CrossRef]
23. Hsu, H.T.; Lee, I.H.; Tsai, H.T.; Chang, H.C.; Shyu, K.K.; Hsu, C.C.; Chang, H.H.; Yeh, T.K.; Chang, C.Y.; Lee, P.L. Evaluate the feasibility of using frontal SSVEP to implement an SSVEP-based BCI in young, elderly and ALS groups. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2015**, *24*, 603–615. [CrossRef] [PubMed]
24. Dias, N.; Mendes, P.; Correia, J. Subject age in P300 BCI. In Proceedings of the 2nd International IEEE EMBS Conference on Neural Engineering, Arlington, VA, USA, 16–19 March 2005; IEEE: New York, NY, USA, 2005; pp. 579–582.
25. Park, D.C.; Reuter-Lorenz, P. The adaptive brain: Aging and neurocognitive scaffolding. *Annu. Rev. Psychol.* **2009**, *60*, 173. [CrossRef]
26. Lotte, F.; Congedo, M.; Lécuyer, A.; Lamarche, F.; Arnaldi, B. A review of classification algorithms for EEG-based brain–computer interfaces. *J. Neural Eng.* **2007**, *4*, R1. [CrossRef] [PubMed]
27. Stancin, I.; Cifrek, M.; Jovic, A. A Review of EEG Signal Features and Their Application in Driver Drowsiness Detection Systems. *Sensors* **2021**, *21*, 3786. [CrossRef] [PubMed]
28. Okahara, Y.; Takano, K.; Komori, T.; Nagao, M.; Iwadate, Y.; Kansaku, K. Operation of a P300-based brain-computer interface by patients with spinocerebellar ataxia. *Clin. Neurophysiol. Pract.* **2017**, *2*, 147–153. [CrossRef]
29. Guy, V.; Soriani, M.H.; Bruno, M.; Papadopoulo, T.; Desnuelle, C.; Clerc, M. Brain computer interface with the P300 speller: Usability for disabled people with amyotrophic lateral sclerosis. *Ann. Phys. Rehabil. Med.* **2018**, *61*, 5–11. [CrossRef]
30. Chen, M.L.; Fu, D.; Boger, J.; Jiang, N. Age-related changes in vibro-tactile EEG response and its implications in BCI applications: A comparison between older and younger populations. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2019**, *27*, 603–610. [CrossRef]
31. Nguyen, P.; Tran, D.; Vo, T.; Huang, X.; Ma, W.; Phung, D. EEG-based age and gender recognition using tensor decomposition and speech features. In Proceedings of the International Conference on Neural Information Processing, Daegu, Republic of Korea, 3–7 November 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 632–639.
32. De Venuto, D.; Annese, V.F.; Mezzina, G. Real-time P300-based BCI in mechatronic control by using a multi-dimensional approach. *IET Softw.* **2018**, *12*, 418–424. [CrossRef]
33. Hashmi, M.F.H.M.F.; Kene, J.D.K.J.D.; Kotambkar, D.M.K.D.M.; Matte, P.M.P.; Keskar, A.G.K.A.G. An efficient and high accuracy P300 detection for brain computer interface system based on kernel principal component analysis. *Preprint* **2021**. . [CrossRef]
34. Kwak, N.S.; Müller, K.R.; Lee, S.W. A lower limb exoskeleton control system based on steady state visual evoked potentials. *J. Neural Eng.* **2015**, *12*, 056009. [CrossRef]
35. Nurseitov, D.; Serekov, A.; Shintemirov, A.; Abibullaev, B. Design and evaluation of a P300-ERP based BCI system for real-time control of a mobile robot. In Proceedings of the 2017 5th International Winter Conference on Brain-Computer Interface (BCI), Gangwon, Republic of Korea, 9–11 January 2017; IEEE: New York, NY, USA, 2017; pp. 115–120.
36. Kaur, B.; Singh, D.; Roy, P.P. Age and gender classification using brain–computer interface. *Neural Comput. Appl.* **2019**, *31*, 5887–5900. [CrossRef]
37. Craik, A.; He, Y.; Contreras-Vidal, J.L. Deep learning for electroencephalogram (EEG) classification tasks: A review. *J. Neural Eng.* **2019**, *16*, 031001. [CrossRef]
38. Komolovaitė, D.; Maskeliūnas, R.; Damaševičius, R. Deep Convolutional Neural Network-Based Visual Stimuli Classification Using Electroencephalography Signals of Healthy and Alzheimer's Disease Subjects. *Life* **2022**, *12*, 374. [CrossRef] [PubMed]
39. Kaushik, P.; Gupta, A.; Roy, P.P.; Dogra, D.P. EEG-based age and gender prediction using deep BLSTM-LSTM network model. *IEEE Sens. J.* **2018**, *19*, 2634–2641. [CrossRef]

40. Khasawneh, N.; Fraiwan, M.; Fraiwan, L. Detection of K-complexes in EEG waveform images using faster R-CNN and deep transfer learning. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 297. [CrossRef]
41. Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S.M.; Hung, C.P.; Lance, B.J. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* **2018**, *15*, 056013. [CrossRef]
42. Dillen, A.; Steckelmacher, D.; Efthymiadis, K.; Langlois, K.; De Beir, A.; Marušič, U.; Vanderborght, B.; Nowé, A.; Meeusen, R.; Ghaffari, F.; et al. Deep learning for biosignal control: Insights from basic to real-time methods with recommendations. *J. Neural Eng.* **2022**, *19*, 011003. [CrossRef] [PubMed]
43. Lopez-Calderon, J.; Luck, S.J. ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Front. Hum. Neurosci.* **2014**, *8*, 213. [CrossRef] [PubMed]
44. Gembler, F.; Stawicki, P.; Rezeika, A.; Volosyak, I. A comparison of cVEP-based BCI-performance between different age groups. In Proceedings of the International Work-Conference on Artificial Neural Networks, Gran Canaria, Spain, 12–14 June 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 394–405.
45. Zhang, X.; Jiang, Y.; Hou, W.; Jiang, N. Age-related differences in the transient and steady state responses to different visual stimuli. *Front. Aging Neurosci.* **2022**, *14*, 1004188. [CrossRef] [PubMed]
46. Volosyak, I.; Gembler, F.; Stawicki, P. Age-related differences in SSVEP-based BCI performance. *Neurocomputing* **2017**, *250*, 57–64. [CrossRef]
47. Nasreddine, Z.S.; Phillips, N.A.; Bédirian, V.; Charbonneau, S.; Whitehead, V.; Collin, I.; Cummings, J.L.; Chertkow, H. The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **2005**, *53*, 695–699. [CrossRef]
48. Potts, G.F. An ERP index of task relevance evaluation of visual stimuli. *Brain Cogn.* **2004**, *56*, 5–13. [CrossRef] [PubMed]
49. Delorme, A.; Makeig, S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **2004**, *134*, 9–21. [CrossRef] [PubMed]
50. MATLAB. *Version 7.10.0 (R2010a)*; The MathWorks Inc.: Natick, MA, USA, 2010.
51. Pion-Tonachini, L.; Kreutz-Delgado, K.; Makeig, S. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage* **2019**, *198*, 181–197. [CrossRef] [PubMed]
52. Luck, S.J.; Gaspelin, N. How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology* **2017**, *54*, 146–157. [CrossRef] [PubMed]
53. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
54. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
55. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794.
56. Ling, C.X.; Huang, J.; Zhang, H.; et al. AUC: A statistically consistent and more discriminating measure than accuracy. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, IJCAI-03, Acapulco, Mexico, 9–15 August 2003; Volume 3, pp. 519–524.
57. Cecotti, H.; Ries, A.J. Best practice for single-trial detection of event-related potentials: Application to brain-computer interfaces. *Int. J. Psychophysiol.* **2017**, *111*, 156–169. [CrossRef]
58. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
59. Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. [CrossRef]
60. Bekkar, M.; Djemaa, H.K.; Alitouche, T.A. Evaluation measures for models assessment over imbalanced data sets. *J. Inf. Eng. Appl.* **2013**, *3*, 10.
61. Saeidi, M.; Karwowski, W.; Farahani, F.V.; Fiok, K.; Taiar, R.; Hancock, P.A.; Al-Juaid, A. Neural decoding of eeg signals with machine learning: A systematic review. *Brain Sci.* **2021**, *11*, 1525. [CrossRef] [PubMed]
62. Bianchi, L.; Sami, S.; Hillebrand, A.; Fawcett, I.P.; Quitadamo, L.R.; Seri, S. Which physiological components are more suitable for visual ERP based brain–computer interface? A preliminary MEG/EEG study. *Brain Topogr.* **2010**, *23*, 180–185. [CrossRef]
63. Volosyak, I.; Rezeika, A.; Benda, M.; Gembler, F.; Stawicki, P. Towards solving of the Illiteracy phenomenon for VEP-based brain-computer interfaces. *Biomed. Phys. Eng. Express* **2020**, *6*, 035034. [CrossRef] [PubMed]