Analysis of Window-Delay Score for Data Augmentation Methods in Brain-Computer Interfaces

João Stephan S. Maurício[1], Marcelo M. Amorim[1], Alex Borges[1], Heder Bernardino[1], Gabriel de Souza[1]

[1] Federal University of Juiz de Fora (UFJF) – Juiz de Fora – MG – Brazil

{joaossmauricio,marcelodmeloamorim}@gmail.com, gabriel.souza@ufjf.br,

{alex.borges,heder}@ice.ufjf.br

Abstract. *Post-stroke motor rehabilitation is a challenging problem in the medical field. Considering this, Brain-Computer Interfaces (BCI) have proven to obtain positive results, especially for chronic stroke. However, as electroencephalogram data collection for BCI can be challenging, Data Augmentation (DA) methods can reduce data collection and simplify training. This study proposes analyzing the temporal behavior of the accuracy instead of analyzing it in fixed intervals, as it is commonly done. Six DA methods and five classification models were evaluated for different scenarios. Results show Filter Bank Common Spatial Pattern is consistent while EEGNet peaks at 2.5 seconds. Sliding Window DA improves response time by 16% and enhances model robustness.*

1. Introduction

Strokes are one of the leading causes of death in the world, they also reduce the quality of life and can lead to paralysis in limbs [Pacheco-Barrios et al. 2022]. Considering this, the Motor Imagery (MI) paradigm in the Brain-Computer Interface (BCI) field have been profusely used in motor rehabilitation. BCI is a way to connect the brain and an external device [Wolpaw et al. 2000]. Commonly, this device is a computer that receives the electrical activity from the brain. Electroencephalogram (EEG) is the most common device for collecting brain signals since it is cheaper, faster, and easier to apply. Besides its practicality, it is safer than the invasive collection of signals, which uses internal electrodes in the brain. Even with it being faster than other methods, collecting enough data for training models can be tedious for the subject.

Data augmentation (DA) is a well-known procedure in Machine Learning. Many methods have been proposed in Computer Vision and Natural Language Processing [Li et al. 2022a, Mumuni and Mumuni 2022], among other areas. Those approaches can also be used in the MI paradigm. In this case, DA methods are growing and showing positive results in the MI-BCI area [Faria et al. 2022], presenting advantages such as avoiding overfitting and increasing the training data when data is scarce [Freer and Yang 2020].

Usually, a fixed time accuracy is used to evaluate the DA methods. That can lead to discarding classifiers based on a punctual result instead of analyzing its full temporal behavior; Considering this, this article proposes to evaluate the temporal behavior of DA methods with different classifiers. The results show that Filter Bank Common Spatial Pattern (FBCSP) has more temporal robustness than EEGNet. Moreover, it is shown that Sliding Window (SW) can increase EEGNet's stability over time.

2. Related Work

The first in-depth comparison of DA methods in the MI-BCI area compared Riemannian Classifier, Convolutional Neural Networks, and Convolutional Long-Short Time Memory (C-LSTM) when using the following DA methods [Freer and Yang 2020]: Gaussian Noise (GN), Signal Multiplication, Signal Flip, and Frequency Shift. The results pointed out that average overall accuracy increases with DA.

Experiments similar to those in [Freer and Yang 2020] were performed in [Zhang et al. 2020]. However, in [Zhang et al. 2020], DA was applied in the Time-Frequency domain after a Short-Time Fourier Transform, with GN, Geometric Transformation, and Deep Convolutional Generative Adversarial Networks (DCGAN). The augmented and non-augmented data were then classified using a Deep Neural Network (DNN) also in the Time-Frequency domain. DCGAN provided better results than traditional DA methods, and DA improved classification.

In a later paper [Lashgari et al. 2021], experiments were performed with five DA methods using a proposed Convolutional Neural Network (CNN) classifier. The five DA methods are Sliding Window (SW), GN, a Generative Adversarial Network (GAN), Time-Frequency Recombination (TFR) and Empirical Mode Decomposition (EMD). The proposed Neural Network (NN) outperforms state-of-the-art with and without DA.

An evaluation of five different DA methods (GN, SW, TFR, Time-Frequency GN, EMD, and SW+TFR) was also performed using EEGNet [Faria et al. 2022]. The experiment was performed by reducing the original data and filling it with DA, obtaining the best result with SW+TFR.

Instead of evaluating BCI for a single paradigm, there was also a comparison of three DA methods (Performance-Measure-Based Time Warp, Time-Frequency GN, and Frequency Masking) for Awareness Recognition, MI, and Steady-State Visually Evoked Potentials (SSVEP) [Li et al. 2022b]. It was concluded that the CNN models were improved when using DA.

A Spatial Variation Generation DA method for MI was proposed by [Qin et al. 2023]. It compared Hemisphere Perturbation, GN, Random Shift, Mixup, Frequency Shift, and TFR and five Deep Learning models, obtaining promising results. Many datasets were used in the related works described, such as BCI III Competition, BCI IV Competition (1, 2a, and 2b), Taiwan Driving Dataset, Sleep Physionet, and Physionet MI. All these datasets differ in the number of subjects, trials, and electrodes, which shows how wide are the situations where DA can be used.

There are also many other which have proposed DA methods in the literature [Fahimi et al. 2021, Kim et al. 2023, Luo et al. 2021] or which have used known methods [Huang et al. 2020, Choi et al. 2022, Yang et al. 2021]. Even though many articles present a robust analysis of the classifiers, DAs, and datasets used, none of them analyzed the accuracy through the trials. Usually, when analyzing accuracy, the common practice is fixing a time and getting the average per trial. However, analyses with a fixed time can be quite a narrow approach to BCI as this is a multi-objective problem. This negligence can lead to choosing models with higher accuracies than others but requires a long delay to reach this quality level. Here, Window-Delay Score [de Souza et al. 2023] is used to broaden the view when analyzing DA methods.

3. Datasets

The BCI Competition IV [Tangermann et al. 2012], as well as the previous competitions, consists of delivering high-quality open-access data for BCI. Also, as it is a competition, the challenge is to see scientists from many areas other than BCI proposing new ideas and enhancing the analysis methods. Usually, these competitions focus on accuracy or Kappa score, guiding models to a biased path. Even so, the datasets still contribute a lot to the evolution of research in this area.

3.1. BCI Competition IV 2a

The dataset 2a of BCI Competition IV (BCICIV2a) was recorded with 22 electrodes with a sampling of 250Hz. The signal was collected between $0.5 - 100$Hz and a notch filter of 50Hz. It was recorded for nine subjects in two sessions, with 288 trials each. The 288 trials are equally divided for each class: left-hand, right-hand, tongue, and feet. Each trial begins with a warning sound, and in the first two seconds, a fixation cross is displayed on the computer screen. After this, a cue is presented in the next 1.25 seconds overlapping with the motor imagery that starts at the third second. After three seconds of imagery, the trial ends with a total duration of 6 seconds, followed by a brief pause until the new trial.

3.2. BCI Competition IV 2b

Dataset 2b of BCI Competition IV (BCICIV2b) has a configuration of 3 electrodes (C3, Cz, C4) also with a sampling of 250Hz. Two filters were applied during the acquisition: a $0.5-100$Hz bandpass filter and a 50Hz notch filter. Its data was recorded for nine subjects in five sessions with 120 trials per session equally divided amongst the two classes: left-hand and right-hand. The first two sessions did not show feedback for the subject, unlike the last three, which show screening feedback. The sessions without feedback begin with a fixation cross shown for 3 seconds with a warning sound in the second second. The cue for the imagery starts right after these three seconds, for 1.25 seconds. The cue overlaps in the last fifth with the beginning of the imagery starting at the second four and lasts three seconds. Following this, there is a pause for the new trial. The other three sessions are different. Firstly, a grey smiley face is presented for 3.5 seconds, followed by a smiley feedback imagery for 4 seconds. Also, a warning sound is played in the second second. The cue is shown during the whole feedback imagery, starting half a second sooner.

4. Classifier Methods

This section presents the models used for the experiments with DA methods. For spatial filtering, Common Spatial Pattern is described, and its complement with a Filter Bank. The Single Electrode Energy model and the Convolutional Neural Network EEGNet are also described.

4.1. Common Spatial Pattern

Common Spatial Pattern (CSP) is a spatial filter that increases the difference of the signals for two classes while reducing the difference inter-class. CSP transforms the data by a linear transformation as $X_i = W^T X_i$, where X_i is the i-th trial of the training data, and W^T is the fitted matrix found out by CSP. W^T is composed by $\frac{m}{2}$ first columns and $\frac{m}{2}$ last columns from the generalized eigenvalue problem of CSP [Ang et al. 2012]. The

transformation matrix of CSP has m columns and E rows, where E is the number of electrodes and m is a hyper-parameter of CSP. The LogPower feature extraction function is commonly used to optimize the CSP performance. In this approach, LogPower is defined as

$$Z_i = log\left(\frac{diag(R_i)}{tr(R_i)}\right) \tag{1}$$

where R_i is the correlation matrix given by $R_i = X_i X_i^T$, $diag(\cdot)$ is the diagonal of a matrix, and $tr(\cdot)$ is the trace of a matrix. Moreover, CSP is also used as a pipeline's name with a bandpass filter and a classifier. In this work, the CSP pipeline is composed of a bandpass filter, CSP filter, LogPower extraction, and the classifier Naive-Bayes Parzen Window (NBPW).

4.2. Filter Bank Common Spatial Pattern

Filter Bank CSP (FBCSP) is a BCI pipeline that uses CSP with many bandpass filters [Ang et al. 2012]. In FBCSP, the training data pass through a set of bandpass filters, followed by CSP in each sub-band. As the number of features extracted using FBCSP is larger than CSP, a filter selection step is added to FBCSP. The complete FBCSP pipeline has a set of bandpass filters, a set of CSP filters, the LogPower function, the feature selection Mutual Information-based Best Individual Feature (MIBIF), and the NBPW classifier.

4.3. Single Electrode Energy

Single Electrode Energy (SEE) [de Souza et al. 2021] is a simple BCI pipeline for MI that uses only a single electrode. This method applies the LogPower function presented in Equation 1 after a bandpass filter, and its feature value is used to classify the signal. Two ways to classify the signal in SEE were proposed, namely: (i) Median-SEE: using the median of the training data to separate the classes; or (ii) Sigmoid-SEE: fitting a sigmoid function using the training data.

4.4. EEGNet

EEGNet [Lawhern et al. 2018] is a CNN based on FBCSP [Ang et al. 2012]. It aims to join all the steps used in FBCSP in its architecture. EEGNet is composed of a temporal convolution set to perform the temporal filtering. The spatial convolution represents the spatial filter from the standard BCI pipeline. After that, it has a separated convolution, similar to feature extraction, and, finally, a Softmax layer performs the classification. CNN topology presented better results when compared to other Deep Learning techniques such as shallow-ConvNet and deep-ConvNet. Moreover, due to its simplicity, EEGNet can be implemented in Field Programmable Gate Arrays (FPGA) or other embarked systems.

5. Data Augmentation Methods

Data augmentation is a technique in which data is amplified with slightly modified copies of its instances [Mumuni and Mumuni 2022]. It usually prevents overfitting in models, but it can also be used where the initial dataset is small or to improve the model's accuracy. Here, we evaluate six data augmentation methods: Gaussian Noise, Sliding

Window, Time-Frequency Recombination (Fixed Time), Time-Frequency Recombination (Fixed Frequency), Time-Frequency Gaussian Noise, and Empirical Mode Decomposition[1]. Some methods can be applied to any dataset, such as Gaussian Noise. On the other hand, other ones are more specific to EEG data, such as Time-Frequency Recombination.

5.1. Gaussian Noise

The Gaussian Noise addition is the simplest data augmentation method evaluated in this work. It consists of perturbing the original data with a normal distribution noise. Then, the perturbed signal is inserted in the original dataset to increase its size. This generation of an artificial trial X_i' can be expressed as $X_i' = X_i + \xi$ where $X_i \in \mathbb{R}^{E \times T}$ is the i-th trial from the original dataset, E is the number of electrodes, T its the number of timestamps in the trial, and $\xi \sim \mathcal{N}(\mu, \sigma^2)$. After that transformation, the artificial trial has the same dimension and domain as the original trial.

5.2. Sliding Window

Sliding Window is a method that uses different window positions by shifting its start time [Faria et al. 2022]. This method uses the features collected by different timestamps to prevent overfitting. When using the Sliding Window method, no artificial trial is created. Instead, the used windows are shifted. Therefore, SW can find out more features due to the non-stationary behavior of EEG signals.

5.3. Time-Frequency Recombination (Fixed Time)

Time-Frequency Recombination-Fixed Time (TFR-T) consists of decomposing the input into different segments and then combining segments from random trials to reconstruct an artificial signal [Lotte 2015]. The following steps describe this method:

1. Each trial is converted to the time-frequency domain using a Short-Time Fourier Transform (STFT).
2. Different trials are drawn and grouped by each segment time, forming an artificial trial M_i' as
$$M_i' = [M_{(1,r)}, M_{(2,r)}, ..., M_{(T',r)}] \tag{2}$$
where $M_i \in \mathbb{C}^{E \times F \times T'}$, the dimension T' represents the segment timestamp and F the frequencies in which the signal is decomposed. Also, in Equation (2), r is a different random number for each segment time in the interval $1, ..., N$, and N is the number of trials of a given class.
3. The new trials $M'^{(i)}$ are converted back to the time domain by applying an inverse STFT.

5.4. Proposed approach Time-Frequency Recombination (Fixed Frequency)

Time-Frequency Recombination-Fixed Frequency (TFR-F) is performed as TFR-T presented in Section 5.3. However, in the second step of the method, the draws are performed by grouping the frequency bins instead of concerning time as $F_i' = [F_{(1,r)}, F_{(2,r)}, ..., F_{(F,r)}]$ where $F_i \in \mathbb{C}^{E \times T' \times F}$. In addition, r is a random number in $1, ..., N$ drawn from a uniform distribution that is different for each frequency in the Time-Frequency domain. Then, the new artificial trials go through the inverse STFT and are inserted into the original dataset.

[1]https://github.com/stephanJoao/bci-data-augmentation

5.5. Time-Frequency Gaussian Noise

Time-Frequency Gaussian Noise (TFGN) works similarly to the Gaussian Noise method presented in Section 5.1. However, similarly to TFR-F, the noise is added in the time-frequency domain. Therefore, it is converted using an STFT as above, creating $TF_i \in \mathbb{C}^{E \times F \times T'}$. After this, the amplitude and phase of the complex numbers in TF_i are extracted into $A^{(i)} \in \mathbb{R}^{E \times F \times T'}$ and $\phi^{(i)} \in \mathbb{R}^{E \times F \times T'}$. To generate a new trial from this, a trial is coped, and its amplitude is perturbed as $A'^{(i)} = A^{(i)} + \xi$ where ξ is a random value sampled from a normal distribution $\xi \sim \mathcal{N}(\mu,\, \sigma^2)$. The artificial trials can be obtained by applying $TF'_i = A'_i \cdot e^{\phi_i j}$ to each trial. After that, the inverse STFT transforms the signal back to the time domain. Finally, the artificial trial is inserted into the original dataset.

5.6. Empirical Mode Decomposition

Empirical Mode Decomposition is a well-known method in signal processing [Huang et al. 1998], in which a nonlinear and non-stationary signal is decomposed into a set of intrinsic mode functions (IMF) and a residue. With this technique, a trial X_i can be decomposed in K IMFs as

$$X_i = \sum_{k=1}^{K} c_{i,k} + r_{i,K} \tag{3}$$

where r_n is the residue and $c_{i,k}$ is the k-th IMF for the i-th trial. Then, different IMFs are randomly chosen and summed to form a new artificial signal.

6. Computational Experiments

This section describes the computational experiments and their results. The training window for the data augmentation methods is $[0.5 - 2.5\text{s}]$ where 0s is the cue onset. The exception was the SW, which varies from $[0 - 2\text{s}]$ to $[2 - 4\text{s}]$ with a step of 0.5s. Thus, SW increases the train data by a factor of 5 while the other methods increase this data by 1.5. We performed the experiments using a 5-fold stratified cross-validation. The beginning of the test windowing varied from -2s to 2s before the end of the motor imagery since 2s is the window's size. The timestep of the windowing in this interval is 0.1s. For both datasets, we used two classes: left-hand and right-hand. The data was resampled with 128Hz, and a $4 - 40\text{Hz}$ bandpass filter was applied.

We performed preliminary experiments and chose the following parameters for the models. For CSP, we used $m = 2$ pairs. In FBCSP, m was also 2 with eight features selected in MIBIF. In addition, FBCSP has its bandpass filter: $[4 - 8\text{Hz}]$, $[8 - 12\text{Hz}]$, $[12 - 16\text{Hz}]$, $[16 - 20\text{Hz}]$, $[20 - 24\text{Hz}]$, $[24 - 28\text{Hz}]$, $[28 - 32\text{Hz}]$, $[32 - 36\text{Hz}]$, and $[36 - 40\text{Hz}]$. In EEGNet, we used eight temporal convolutions of size 64, two spatial convolutions, and 16 separable convolutions of size 16. The dropout rate for the method was 0.5, a learning rate of 0.001 with 1000 iterations, and a batch size of 64. Its important to highlight that all DA methods were applied only to the training data. For the GN data augmentation, the parameters for the normal distribution are $\mu = 0$ and $\sigma = 0.1$, as in the literature [Lashgari et al. 2021, Faria et al. 2022]. For STFT on TFR-T, TFR-F, and TFGN methods, the size of the STFT window is 128 with the Hann window type. The Gaussian noise added in TFGN uses the same parameters as the normal distribution in GN. Lastly, for EMD, the only parameter used is the maximum number of IMFs as 9.

Table 1. WD-score and accuracy results for 22 electrodes case. The non-dominated results are in boldface.

	Median-SEE	Sigmoid-SEE	CSP	FBCSP	EEGNet
Baseline	(2.4, 0.5475)	(2.9, 0.5262)	(2.5, 0.7473)	**(2.4, 0.8252)**	(2.5, 0.8306)
GN	(2.5, 0.5401)	(3.0, 0.5251)	(2.4, 0.7431)	(2.4, 0.8148)	(2.5, 0.8256)
SW	(2.9, 0.5455)	(2.2, 0.5378)	(2.9, 0.7396)	**(2.2, 0.8171)**	(2.5, 0.7693)
TFR-T	(2.5, 0.5313)	(2.5, 0.5285)	(2.3, 0.7438)	**(2.5, 0.8349)**	(2.5, 0.8148)
TFR-F	(5.1, 0.5112)	**(2.1, 0.5409)**	(2.5, 0.7473)	(2.6, 0.8202)	(2.5, 0.8306)
TFGN	(2.8, 0.5370)	(2.8, 0.5247)	(2.3, 0.7392)	**(2.3, 0.8210)**	(2.5, 0.8218)
EMD	(2.5, 0.5289)	(2.6, 0.5278)	(2.4, 0.7087)	(2.2, 0.6659)	(2.5, 0.8002)

Table 2. WD-score and accuracy results for 3 electrodes case without feedback. The non-dominated results are in boldface.

	Median-SEE	Sigmoid-SEE	CSP	FBCSP	EEGNet
Baseline	(2.4, 0.5295)	(2.3, 0.5326)	(2.6, 0.6076)	(2.5, 0.6487)	(2.5, 0.6321)
GN	(2.4, 0.5473)	**(1.6, 0.5129)**	(2.6, 0.6121)	(2.5, 0.6321)	(2.5, 0.6281)
SW	**(1.7, 0.5237)**	(2.0, 0.5308)	(2.5, 0.6165)	(2.7, 0.6504)	**(2.1, 0.5804)**
TFR-T	(2.4, 0.5313)	(2.1, 0.5188)	(2.5, 0.6103)	**(2.5, 0.6611)**	(2.5, 0.6138)
TFR-F	(2.3, 0.5295)	(2.9, 0.5429)	(2.6, 0.6080)	(2.7, 0.6424)	(2.5, 0.6330)
TFGN	**(1.8, 0.5362)**	(2.3, 0.5321)	(2.5, 0.6000)	**(2.4, 0.6272)**	(2.5, 0.6170)
EMD	(2.3, 0.5424)	(3.4, 0.5286)	(2.7, 0.6138)	(2.5, 0.6518)	(2.5, 0.6027)

Table 3. WD-score and accuracy results for 3 electrodes case with feedback. The non-dominated results are in boldface.

	Median-SEE	Sigmoid-SEE	CSP	FBCSP	EEGNet
Baseline	(2.5, 0.5778)	(2.1, 0.5766)	(2.5, 0.7290)	**(2.6, 0.7701)**	**(2.5, 0.7491)**
GN	(2.5, 0.5792)	(2.4, 0.5776)	(2.5, 0.7210)	(2.9, 0.7682)	(2.5, 0.7386)
SW	(3.7, 0.5657)	(2.7, 0.5624)	(3.0, 0.7306)	**(2.9, 0.7836)**	(2.5, 0.7012)
TFR-T	**(2.3, 0.5857)**	(2.5, 0.5755)	(2.5, 0.7357)	(2.8, 0.7787)	**(2.5, 0.7491)**
TFR-F	(2.5, 0.5624)	(2.4, 0.5722)	(2.7, 0.7336)	**(2.7, 0.7804)**	(2.5, 0.7360)
TFGN	**(2.1, 0.5636)**	(2.4, 0.5650)	(2.8, 0.7224)	(2.8, 0.7720)	(2.5, 0.7386)
EMD	**(2.4, 0.5883)**	(2.5, 0.5729)	(2.5, 0.7255)	(2.8, 0.7771)	(2.5, 0.7107)

We compared the WD-score and accuracy values reached by the methods tested here for each dataset. Moreover, we present the non-dominated results, and, in Section 6.4, compare the average accuracies through the time of trials. The cases analyzed are 22 electrodes, three electrodes without feedback, and three electrodes with feedback.

6.1. Many electrodes

The BCICIV2a dataset was chosen for its better spatial resolution once it has 22 electrodes. With the higher spatial resolution, we need more trials to reduce overfitting, making it a good candidate for DA. We used the left-hand and right-hand classes from BCICIV2a. Table 1 presents the WD-score and accuracy results. FBCSP with TFR-T presents the highest accuracy (0.8349). In this and the following analysis, we will not evaluate the results obtained with median- and sigmoid-SEE methods as they did not give prominent results for any case. Considering this, SW-FBCSP had the lower WD-score, which means it can reach higher accuracy faster than other combinations. It is also clear that the WD-score did not vary significantly from its expected value of $2.5s$, especially for EEGNet. The general results pointed out that the accuracy did not change significantly with the DA methods. The only exception is in FBCSP, in which EMD worsens the accuracy. Figure 1

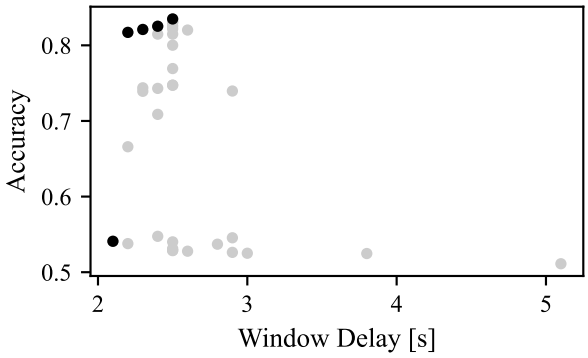


Figure 1. Pareto set for 22 electrodes. Each dot is a pair [DA, Classifier]. Black dots are non-dominated pairs.

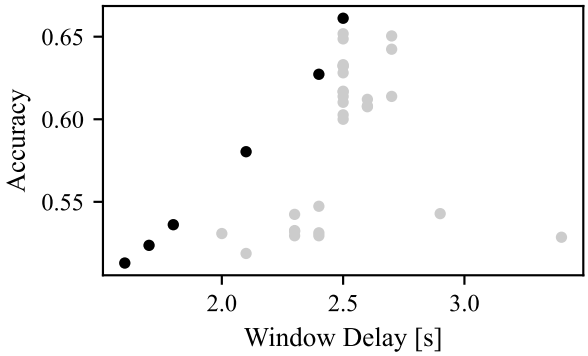


Figure 2. Pareto set for 3 electrodes without feedback. Each dot is a pair [DA, Classifier]. Black dots are non-dominated pairs.

is the Pareto set of the experiment with many electrodes. All non-dominated points are in the interval $2 - 2.5s$, which is expected since $2.5s$ is the end of the training window. All these points are also from FBCSP, with the best result at 2.5. Better results for FBCSP in the many electrode cases are expected due to the spatial characteristics of the method.

6.2. Few electrodes

For few electrodes, we performed the experiments individually for each session. This way, each session of dataset BCICIV2b was considered an individual dataset with a small number of trials. In this section, the results are the junction of the output of the first two sessions, with no feedback to the subject. This experiment aims to evaluate the impact of DA methods when compared to a much larger dataset such as BCICIV2a. Table 2 presents the obtained results. The best result is also obtained by FBCSP with TFR-F with an accuracy of 0.6611. SW-EEGNet found the lowest WD-score Moreover, it was the only DA that decreased the WD-score for EEGNet. Apart from this, the WD-score maintains itself close to the value of $2.5s$ for all methods. Once again, as seen in Figure 2 all non-dominated points are in the interval [$2 - 2.5$s]. The higher ones are both from FBCSP, which is not significantly affected by the DA methods, whereas the point in $2.1s$ is from EEGNet with SW.

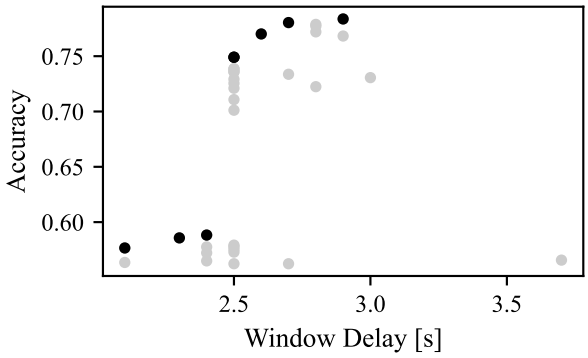


Figure 3. Pareto set for 3 electrodes with feedback. Each dot is a pair [DA, Classifier]. Black dots are non-dominated pairs.

6.3. Few electrodes with feedback

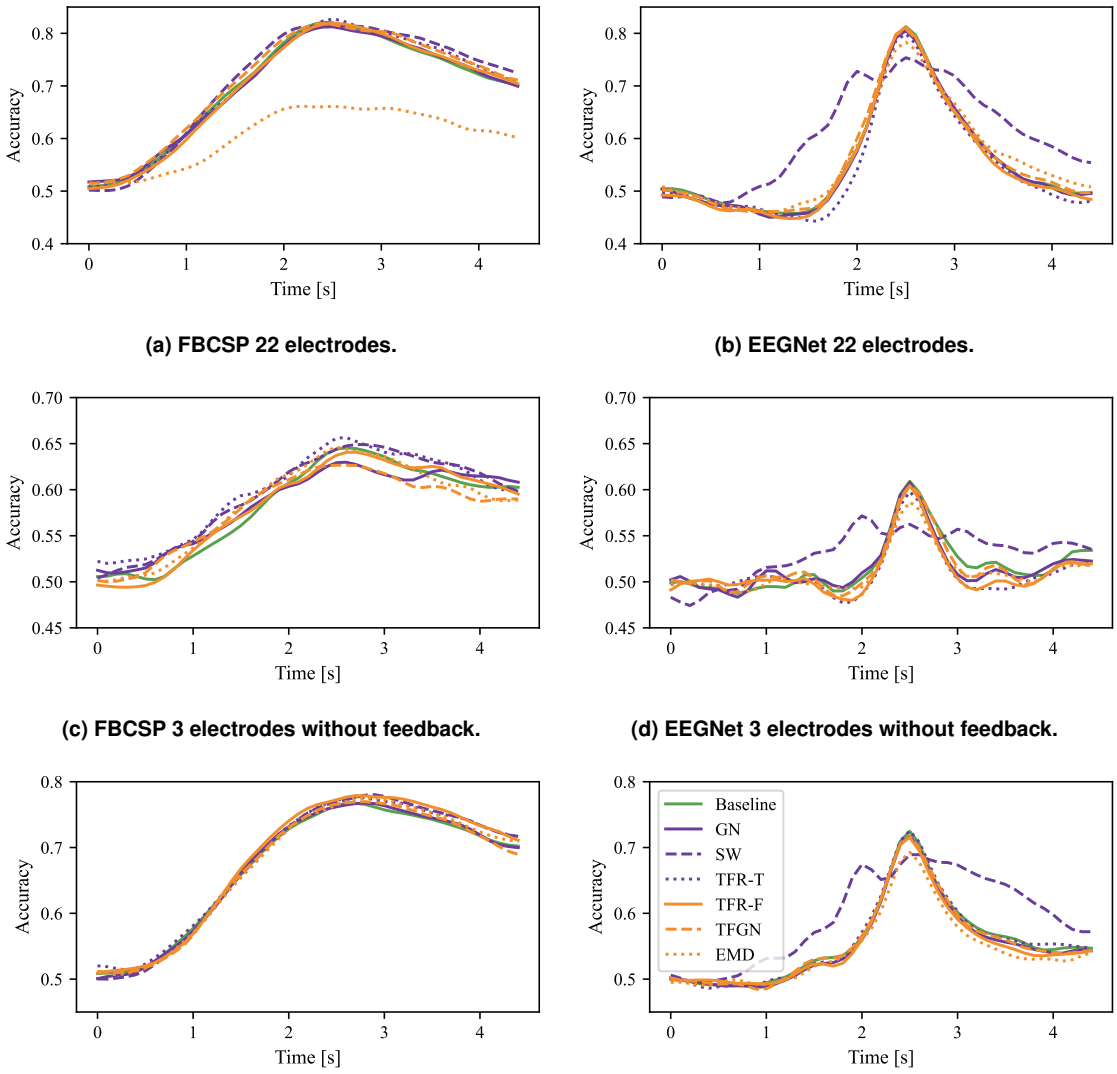
Table 3 presents the WD-score and accuracy results for the sessions with feedback. SW-FBCSP obtained the higher accuracy (0.7836), while EEGNet had the best WD-score. Moreover, EEGNet had the same WD-score no matter the DA method. As seen in Figure 3, the three non-dominated points in the Pareto set from FBCSP are between 2.5 and $3.0s$. WD-score for the case with feedback was higher, as expected given that the subject has its attention divided while waiting for the online feedback.

6.4. Discussion

When using WD-score, we get a fairer accuracy for each combination of model and classifier, but some temporal behavior of the metric is still lost, such as the consistency throughout time. In Figure 4, the average accuracy in the results was plotted for each discrete time for each DA method. We decided to show only FBCSP and EEGNet graphics as they bring more to the table for discussion. In Figures 4a, 4c, and 4e, the DA methods do not influence significantly the accuracy over time with FBCSP, except with EMD in BCICIV2a. Also, it is clear how the accuracy increases smoothly until $2.5s$ and then smoothly decreases. Especially in Figure 4c, DA methods brought more differentiation since the datasets have fewer trials. And, in Figure 4f with feedback, accuracy decreases slower after the peak. Results for the EEGNet can be seen in Figures 4b, 4d, and 4f. EEGNet curve is much steeper than FBCSP around the $2.5s$ peak. For EEGNet, datasets with fewer trials also bring more differentiation, as seen in Figure 4d. In these graphics, even though the other DA methods do not influence much in the accuracy behavior, the SW method increases EEGNet stability over time. This way, the EEGNet curve starts increasing after and takes more time to decrease, widening the peak's plateau. This stability improves the application in real-time since it is desired that the task is executed for $4s$ in the evaluated datasets.

7. Conclusion

Brain-Computer Interfaces have a significant role in post-stroke motor rehabilitation and controlling mechanical prostheses. However, it has some downpoints, such as the tediousness of recording training data and the subject-dependence of the models. Deep learning



(a) FBCSP 22 electrodes.

(b) EEGNet 22 electrodes.

(c) FBCSP 3 electrodes without feedback.

(d) EEGNet 3 electrodes without feedback.

(e) FBCSP 3 electrodes with feedback.

(f) EEGNet 3 electrodes with feedback.

Figure 4. Data augmentation average accuracies through the time of trials.

models obtain positive results in the DA literature. But even with positive results, some studies are narrowly focused on accuracy and Kappa score in just one fixed timestamp.

We analyze here the temporal behavior of accuracy for different classifiers and data augmentation methods using WD-score. Our studies tested five different classifiers: Median-SEE, Sigmoid-SEE, CSP, FBCSP, and EEGNet. Moreover, we used three different datasets: many electrodes with many trials, few electrodes and trials without feedback, and with feedback. For all these cases, we used six different DA methods: GN, SW, two variations of TFR, TFGN, and EMD. All these analyses allowed us to evaluate how DA methods impact the temporal behavior of the accuracy metric. Amongst all cases, few electrodes without feedback had the least number of trials and had more variations in their results. However, the DA methods haven't influenced expressive results or increases, except for SW. The SW method has shown its importance to EEGNet because it increased the stability of EEGNet across time. Furthermore, SW reduced the time for peak accuracy

with EEGNet in 16% for few electrodes without feedback.

For future works, many possibilities arise in the temporal analysis of BCI models. For instance, a combined analysis of Transfer Learning and DA can create a more robust model. Moreover, the evaluation of WD-score for deeper CNN can be performed to verify if the SW behavior remains the same for different architectures. Finally, more DA models can be tested across time or used with other paradigms.

Acknowledgements

We thank the support provided by CAPES, CNPq, FAPEMIG, UFJF, and OpenBCI.

References

Ang, K. K., Chin, Z. Y., Wang, C., Guan, C., and Zhang, H. (2012). Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Frontiers in Neuroscience*.

Choi, H., Park, J., and Yang, Y.-M. (2022). A novel quick-response eigenface analysis scheme for brain–computer interfaces. *Sensors*.

de Souza, G. H., Bernardino, H. S., and Vieira, A. B. (2021). Single electrode energy on clinical brain–computer interface challenge. *Biomedical Signal Processing and Control*.

de Souza, G. H., dos Santos, D. E., Bernardino, H., Vieira, A. B., and Motta, L. P. (2023). Window-delay analysis on eegnet. In *Proceeding of 2023 10th International Conference on Soft Computing & Machine Intelligence*.

Fahimi, F., Dosen, S., Ang, K. K., Mrachacz-Kersting, N., and Guan, C. (2021). Generative adversarial networks-based data augmentation for brain-computer interface. *IEEE Transactions on Neural Networks and Learning Systems*.

Faria, G., De Souza, G. H., Bernardino, H., Motta, L., and Vieira, A. (2022). Analyzing data augmentation methods for convolutional neural network-based brain-computer interfaces. In *Proceedings of the International Joint Conference on Neural Networks*.

Freer, D. and Yang, G.-Z. (2020). Data augmentation for self-paced motor imagery classification with c-lstm. *Journal of Neural Engineering*.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Snin, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and the hubert spectrum for nonlinear and non-stationary time series analysis. In *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*.

Huang, W., Wang, L., Yan, Z., and Liu, Y. (2020). Classify motor imagery by a novel cnn with data augmentation. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*.

Kim, S.-J., Lee, D.-H., and Choi, Y.-W. (2023). Cropcat: Data augmentation for smoothing the feature distribution of eeg signals. In *Proceedings of the International Winter Conference on Brain-Computer Interface, BCI*.

Lashgari, E., Ott, J., Connelly, A., Baldi, P., and Maoz, U. (2021). An end-to-end cnn with attentional mechanism applied to raw eeg in a bci classification task. *Journal of Neural Engineering*.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*.

Li, B., Hou, Y., and Che, W. (2022a). Data augmentation approaches in natural language processing: A survey. *AI Open*.

Li, R., Wang, L., Suganthan, P., and Sourina, O. (2022b). Sample-based data augmentation based on electroencephalogram intrinsic characteristics. *IEEE Journal of Biomedical and Health Informatics*.

Lotte, F. (2015). Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain-computer interfaces. *Proceedings of the IEEE*.

Luo, J., Wang, Y., Xu, R., Liu, G., Wang, X., and Gong, Y. (2021). Channel drop out: A simple way to prevent cnn from overfitting in motor imagery based bci. *Communications in Computer and Information Science*.

Mumuni, A. and Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*.

Pacheco-Barrios, K., Giannoni-Luza, S., Navarro-Flores, A., Rebello-Sanchez, I., Parente, J., Balbuena, A., de Melo, P. S., Otiniano-Sifuentes, R., Rivera-Torrejón, O., Abanto, C., Alva-Diaz, C., Musolino, P. L., and Fregni, F. (2022). Burden of stroke and population-attributable fractions of risk factors in latin america and the caribbean. *Journal of the American Heart Association*.

Qin, C., Yang, R., Huang, M., Liu, W., and Wang, Z. (2023). Spatial variation generation algorithm for motor imagery data augmentation: Increasing the density of sample vicinity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Tangermann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K., Mueller-Putz, G., Nolte, G., Pfurtscheller, G., Preissl, H., Schalk, G., Schlögl, A., Vidaurre, C., Waldert, S., and Blankertz, B. (2012). Review of the bci competition iv. *Frontiers in Neuroscience*.

Wolpaw, J., Birbaumer, N., Heetderks, W., McFarland, D., Peckham, P., Schalk, G., Donchin, E., Quatrano, L., Robinson, C., and Vaughan, T. (2000). Brain-computer interface technology: a review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering*.

Yang, L., Song, Y., Ma, K., and Xie, L. (2021). Motor imagery eeg decoding method based on a discriminative feature learning strategy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Zhang, K., Xu, G., Han, Z., Ma, K., Zheng, X., Chen, L., Duan, N., and Zhang, S. (2020). Data augmentation for motor imagery signal classification based on a hybrid neural network. *Sensors (Switzerland)*.