Data augmentation for learning predictive models on EEG: a systematic comparison

Joseph Paillard[†], **Cédric Rommel**[†], **Thomas Moreau** & **Alexandre Gramfort**

Université Paris-Saclay, Inria, CEA, Palaiseau, 91120, France

E-mail: {`firstname.lastname`}`@inria.fr`

Abstract. The use of deep learning for electroencephalography (EEG) classification tasks has been rapidly growing in the last years, yet its application has been limited by the relatively small size of EEG datasets. Data augmentation, which consists in artificially increasing the size of the dataset during training, has been a key ingredient to obtain state-of-the-art performances across applications such as computer vision or speech. While a few augmentation transformations for EEG data have been proposed in the literature, their positive impact on performance across tasks remains elusive. In this work, we propose a unified and quasi-exhaustive analysis of existing EEG augmentations, which are compared in a common experimental setting. Our results highlight the best data augmentations to consider for sleep stage classification and motor imagery brain computer interfaces, showing predictive power improvements greater than 10% in some cases.

Keywords: Data augmentation, sleep stage classification, brain-computer interface
Submitted to: *J. Neural Eng.*

[†] equal contribution

1. Introduction

Decoding the brain electrical activity is a great scientific challenge for both clinicians and researchers seeking a better understanding of brain dynamics. Recent attempts to leverage deep learning for this difficult task have shown promising results [1]. These new methods have led to performance gains in a wide range of clinically relevant tasks, such as the automatic sleep stage classification from polysomnographic recordings [2, 3, 4, 5]. The ambition of deep learning models is to automatically learn relevant representations from high dimensional data such as EEG [6, 7], while previously used brain decoding methods relied on prior knowledge and handcrafted features [8]. Doing so, deep learning approaches require a less sharp understanding of the underlying neurophysiology and are thus more versatile. Yet, this comes at the cost of using very large training datasets.

Indeed, most of the breakthroughs in deep learning have been enabled by large datasets such as *ImageNet* [9]. Unfortunately, similar datasets do not exist in neuroscience as labeled brain data remains comparatively scarce. Labelling EEG recordings requires a high expertise, is time consuming and can sometimes be inaccurate due to the bias introduced by the human annotator [10]. A second obstacle to the application of deep learning in neuroscience is the high inter-subject variability that is inherent in brain signals [11]. Along with the lack of data, this property makes the generalization on unseen subjects particularly difficult. Without proper regularization, both of these problems can hinder generalization performance and lead to overfitting.

To mitigate the small scale of the neuroscience databases, a promising direction is the use of data augmentation [12, 13, 14]. Data augmentation allows to increase artificially the size of the training set by adding new synthetic examples. These examples are generated by randomly transforming existing ones in a label-preserving way. Doing so, data augmentation helps to teach the decision function to become invariant to the transformation enforced, thus softly reducing the hypothesis space of the training problem. Consequently, it can be interpreted as a regularization method that induces a useful bias by preventing the model from focusing on irrelevant features [15], which in the end makes it less prone to overfitting [16].

Although data augmentation is a well-established method in computer vision it is still under-explored for brain data. Among the few studies using data augmentations for EEG signals, only a fraction of existing transformations is studied simultaneously [17, 18, 19, 20]. Moreover, existing review papers mainly focus on grouping and summarizing results from previous articles instead of carrying out new experiments in comparable settings [1]. In this paper, we propose a unified and quasi-exhaustive analysis of existing data augmentation methods for EEG signals previously proposed in the literature. After presenting our experimental setting and protocol (section 2), we describe the considered transformations that can be used for EEG signals and give insights on the underlying assumptions they make. Then, we assess the effects of these augmentations on standard EEG classification tasks such as sleep stage classification and brain-computer interfaces (BCI) [11]. Finally, we also highlight the important fact that optimal augmentations are not the same for these tasks and also that certain transformations are more useful for certain classes.

In an attempt to organize the plethora of augmentations studied in this manuscript, they are presented in three different groups: augmentations acting on the frequency domain (section 3), augmentations acting on the time domain (section 4) and augmentations acting on the spatial domain (section 5).

2. Experimental protocol

In this section, we present the common experimental protocols used to study each data augmentation on EEG signals. To provide answers in a broad scope, experiments are repeated on two different tasks, sleep stage classification and motor imagery classification in the context of BCI. Both tasks are explained in Section 2.1. Our objective is three-fold: (i) to evaluate the impact of the strength of each transformation (Section 2.2.1), (ii) to compare the global relative benefit of each augmentation depending on the train set size (Section 2.2.2), and (iii) to highlight how the effects of augmentations vary across classes within the same dataset (Section 2.2.3).

2.1. EEG classification tasks

2.1.1. Sleep stage classification

Dataset and preprocessing First experiments were carried out in the context of sleep stage classification. This task is usually performed by sleep experts and is essential to diagnose sleep disorders such as sleep apnea or insomnia. It consists in the classification of 30-second EEG windows into 5 stages following the *American Academy of Sleep Medecine* (*AASM*) manual [21]: Wake (W), Rapid Eye Mouvement (REM) and Non REM stages 1, 2 and 3 (respectively N1, N2 and N3). For this purpose, we used the *SleepPhysionet* dataset [22], which contains whole night polysomnographic recordings from 78 healthy subjects using two EEG channels: Fpz-Cz and Pz-Oz. In this dataset, each signal's window has been annotated by well-trained technicians according to the *Rechtschaffen and Kales* manual [23], before being re-asigned to the more recent stages from the *AASM* manual. As suggested in [5], a minimal data preprocessing is carried, consisting in a lowpass filter with a cutoff frequency of 30 Hz, followed by a simple standardization step (each channel's signal is centered and scaled to have unit variance).

Model and training details The model that is used is a deep convolutional neural network that has been designed for sleep stage classification tasks [5]. It is trained using the Adam optimizer [24] with a learning rate of 10^{-3}. The weighed cross-entropy loss is used to take into account the classes imbalance of the dataset. The batch size is set to 16 to preserve the stochasticity of the gradient descent in very low data regimes. We train the model for 300 epochs using early-stopping with a patience of 30 epochs [5]. Also note that for all our experiments, when using data augmentation, each example in a batch has a probability $p_{\text{aug}} = 0.5$ of being transformed by the augmentation and probability $1 - p_{\text{aug}}$ of being left unchanged.

Splitting strategy EEG recordings have a strong inter-subject variability [11]. Subsequently, to test the trained models in real conditions, some subjects must be set aside and used only for testing in order to avoid subject related information leakage [1]. To take this into account, the following splitting strategy has been defined. First, the dataset is separated into k-folds, each of them containing different subjects. One fold is left out for testing and among the remaining $(k-1)$, 20% of subjects are used for validation. Finally, a subset of the leftover dataset is extracted using a stratified split and used for training. This last step allows to test the model in low data regime while maintaining the distribution of classes.

2.1.2. BCI

Dataset and preprocessing Likewise, experiments are carried out with the *BCI IV 2a* dataset [25]. It consists of recordings from 9 subjects using 22 EEG electrodes. The subjects were asked to perform four motor imagery tasks, namely to imagine the movement of the left hand (class 1), right hand (class 2), both feet (class 3), and tongue (class 4). This dataset was preprocessed using a band-pass filter between 4 Hz and 38 Hz followed by an exponential moving standardization as in [26]. Then trials of 4.5 seconds are used as inputs. Each trial starts 0.5 seconds before the cue that tells the subject to perform the motor imagery task and ends when the cue disappears.

Model and training details The model used is a generic deep convolutional network [26], as implemented in the library BRAINDECODE [27]. It can take advantage of the spatio-temporal structure of the data using spatial filters and convolutions across time. The architecture is inspired by the success of Common Spatial Patterns (CSP) methods [8]. Following the work of [26], the gradient descent is made using the AdamW optimizer [28] with a learning rate of 6.25×10^{-4}. The model is trained for 1600 epochs using early stopping with a patience of 160 epochs and a batch size of 64 [26]. As for sleep staging experiments, each example in a batch has a probability $p_{\text{aug}} = 0.5$ of being transformed by the augmentation.

Splitting strategy According to the rules of the BCI competition [25], for each subject, our model is trained on the first session (or a fraction of it) and evaluated on the second session. The experiment is repeated across all nine subjects.

2.2. Experiments

In this section, we describe the three types of experiments carried out with each EEG augmentation considered.

2.2.1. Parameters selection To address the implications of the strength of the transformation, it should first be acknowledged that several augmentations have a parameter that can be adjusted to control how strongly the inputs are transformed. For example, the Gaussian noise augmentation has a parameter σ corresponding to the standard deviation of the distribution from which the noise is sampled. The choice of the value of such a parameter is thus crucial and even as important as the choice of the augmentation itself, as later depicted in our results (sections 3, 4 and 5).

This experiment unfolds in two steps: 1) narrowing down the range of parameter values and 2) carrying out a grid-search. For the first step, an upstream manual exploration allows to estimate an interval with an upper bound above which the augmentation distorts too much the relevant information contained in the signal. In the case of Gaussian noise, with σ values greater than 0.2, EEG signals become so noisy that the augmentation is systematically detrimental to the learning. For the second step, a grid-search is carried out using 11 linearly spaced values within the aforementioned interval. For each parameter value, a 10-fold cross validation score is computed using the balanced accuracy metric. Since data augmentation is all the more efficient in low data regimes (as shown in our experiments from Sections 3.2.2, 4.2.2 and 5.2.2), we carried our parameter selection using a small balanced fraction of the initial datasets (e.g. 2^{-7} for *SleepPhysionet* dataset) to make potential improvements more apparent.

2.2.2. Learning curves The second experiment aims at comparing the benefits brought by different augmentation methods. To this end, for each augmentation operation, we compute a learning curve which shows the model's performance when it is trained on fractions of the training set. The results obtained with each data augmentation are then compared to a baseline, consisting in the same model trained with no data augmentation.

2.2.3. Per class analysis Finally, to get a deeper understanding of the effects of data augmentations, we take a closer look at single points from the learning curve and analyze how effects vary depending on the class. This observation is guided by the intuition that the invariances encoded by data augmentations might be more relevant for some classes than others. For example, the channel symmetry augmentation, which switches EEG channels from left and right hemispheres, is much more relevant for non-lateralized brain activities such as imagining tong movements, whereas it is likely detrimental for lateralized functions, such as right- or left-hand movements.

The first experiments on the *SleepPhysionet* dataset reveals that augmentations are systematically more helpful in low data regimes (*cf.* Section 3.2.2). Hence, data augmentation has a greater effect on underrepresented classes. Since we are seeking to assess the effects of transformations on learned representations for each class, these experiments require class proportions to be equalized before training. To do this, a subsampling step is added to the pre-processing pipeline, allowing to work with balanced data.

3. Frequency domain augmentations

3.1. Rationale of the transformations

Frequency shift For many EEG classification tasks, it is believed a substantial part of the information lies in the frequency domain of the recording. In sleep scoring for example, most sleep stages are characterized by the occurrence of specific brain rhythms in a given frequency range. The stage N2 is for instance characterized by so-called sleep spindles with frequencies located between 12 and 15 Hz, while stage N3 is characterized by slow waves in the delta band [21].

The predominance of certain rhythms is well captured by the power spectrum density (PSD) of a signal, which shows peaks as depicted in Figure 1. The `FrequencyShift` augmentation, proposed in [29], shifts the PSD of the signal by a factor Δf. The rationale behind this augmentation is that EEG recordings have a strong inter-subject variability. Consequently, the locations of frequency bands with a high power density are likely to be slightly different between two subjects even though they correspond to the same category of cerebral activity, as shown in Figure 1.

Since EEG recordings produce real valued signals, their Fourier transform has Hermitian symmetry, which means that $\mathcal{F}[X](f) = \mathcal{F}[X]^*(-f)$, where $\mathcal{F}[X]^*$ denotes the complex conjugate of $\mathcal{F}[X]$. As a result, shifting the Fourier transform of the signal towards high or low frequencies would break this symmetry. To avoid this, the shift is performed on the complex analytic signal associated to X, $X_a = X + j\mathcal{H}(X)$, where $\mathcal{H}$ denotes the Hilbert transform. Then taking the real part allows to recover the shifted signal:

$$\texttt{FrequencyShift}[X](t) := \mathrm{Re}(X_a(t) \cdot e^{2i\pi\Delta f \cdot t}).$$

The strength of this transformation can be set through the parameter Δf that controls the shift of the PSD. This parameter is randomly sampled with uniform probability in an interval [-Δf_{max}, $+\Delta f_{max}$] each time an EEG window is augmented.

This augmentation moderately modifies the representation of the signal in the frequency domain. Hence, it also changes globally its representation in the time domain. As seen in Figure 1, by associating the same label to signals that have a slightly shifted PSD, this augmentation encourages the model to become invariant to small inter-subject PSD peak frequency variations.

Fourier transform surrogate As just mentioned, the information contained in the PSD is instrumental in EEG classification tasks. Taking this into account, one may wonder how to transform a recording in a way that leaves the PSD unchanged while generating a different waveform. This is precisely what is achieved with the Fourier Transform surrogates method [17]. This augmentation generates so-called surrogates by randomizing the phase of the Fourier coefficients of a recording, while leaving their amplitudes unchanged.

More concretely, this is performed by adding random noise to the phase of the Fourier coefficients of a signal, followed by the application of the inverse Fourier transform:

$$\mathcal{F}[\texttt{FTSurrogate}(X)](f) = \mathcal{F}[X](f)e^{i\Delta\varphi},$$

where $\mathcal{F}$ is the Fourier transform operator, f is a frequency and $\Delta\varphi$ is a frequency specific random phase perturbation. In our implementation, a value of $\Delta\varphi$ is uniformly sampled in an interval $[0, \Delta\varphi_{\mathrm{max}}]$ for each frequency f, where $\Delta\varphi_{\mathrm{max}} \in [0, 2\pi)$ is a hyperparameter.

While this augmentation preserves the frequency-bands power ratios, it triggers a global change of the signal's representation in the time-domain. As a result, it decreases the model's reliance on the signal's time representation, encouraging the model to base its predictions only on the PSD. Thus, a model trained with `FTSurrogate` is expected to have a higher misclassification rate on sleep stages such as N2, which are strongly characterized by specific patterns in the time-domain. This is illustrated on Figure 2, which shows that characteristic patterns such as K-complexes are erased by this operation.

The Fourier transform surrogate method is based on the assumption that EEG signals are generated by stationary linear random processes [17]. As such, they must be uniquely described by the amplitudes of their Fourier coefficients and must have random phases in $[0, 2\pi)$. It also considers each frequency as independent, which might be a too strong assumption. Indeed, neural signals contain transient events, such as K-complexes, which spread across multiple frequencies, hence introducing some dependence among the phases of multiple Fourier coefficients (cf. Figure 2). Besides, phenomena known as cross-frequency coupling (CFC) have been reported in human electrophysiology signals [30]. Nevertheless, the hypothesis of weakly stationary Gaussian linear stochastic

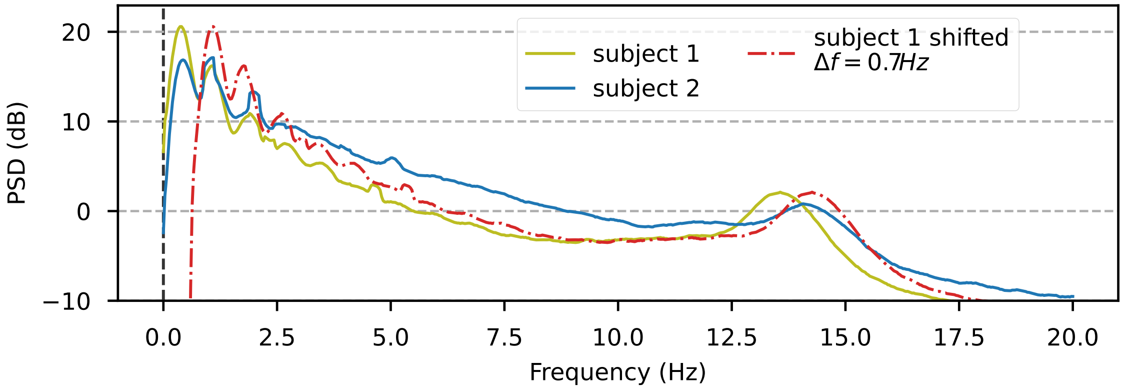


Figure 1: Averaged power spectrum density of windows corresponding to the sleep stage N2, for two different subjects from the *SleepPhysionet* dataset. The red dash-dot curve corresponds to the recording of subject 1 transformed using the `FrequencyShift` augmentation. It allows to translate the PSD peak close to subject 2.

processes have been shown to be compatible with EEG signals [31]. Plausible arguments used to advocate it are the huge number of neurons included in an EEG recording, the complicated structure of the brain and the possible blur of dynamical structures due to the different conductivities of the skull and other intermediate tissues. The consequence of this argument is that, while the Fourier transform surrogate method might be useful on the context of EEG, it is likely to be less relevant for intracranial recordings.

Band-stop filter To the same extent, the band-stop filter augmentation transforms the signal in the frequency domain by filtering out a given frequency band. This augmentation aims to prevent machine learning models from overfitting on subject specific features and from relying too much on a few narrow frequency regions [18, 32].

The implementation of this augmentation uses the finite impulse response notch filter from *MNE-Python* [33]. For each augmented EEG signal, the center of the frequency band is randomly picked from a uniform distribution between 0 and 38 Hz, which is the approximate low-pass filtering frequency used to preprocess both datasets. The width of the filter is a parameter that must be adjusted by the user.

By removing random frequency bands, this transformation can be compared to a dropout layer [16] that prevents the model from relying on specific frequencies. Hence, this transformation enforces the invariance of the decision function to the disappearance of certain frequency bands. Additionally, since this transformation acts on the Fourier transform of the signal, it globally affects the time domain representation of the EEG recordings.

3.2. Experimental results

3.2.1. Parameters selection The parameters to be set for frequency domain augmentations are listed in Table 1.

SleepPhysionet The results on the sleep staging task presented in Figure 3 and Table 1, reveals that the optimal strength varies signif-

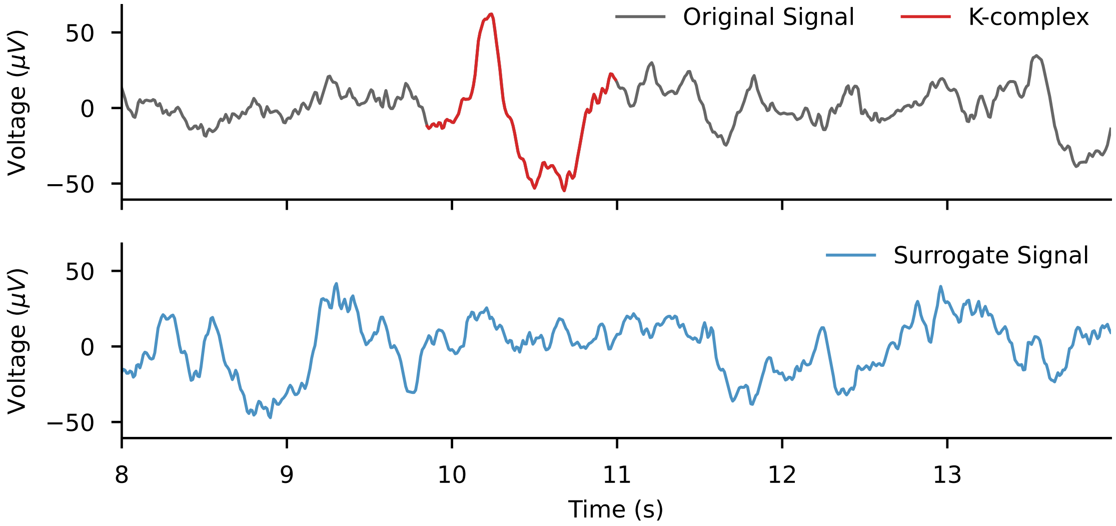


Figure 2: Effect of `FTSurrogate` on transient patterns. The original extract from a window scored as N2 presents a highly localized K-complex, whereas the surrogate does not.

icantly from one transformation to the other. The `FTSurrogate` augmentation benefits from the larger possible range of $\Delta\varphi$ values, corresponding to a nearly completely random phase selection within the interval $[0, 2\pi)$. On the contrary, `FrequencyShift` works better for small Δf values. An interpretation for this result could be that small frequency shifts allow to capture the inter-subject variability whereas larger values mix-up the frequency bands characterizing different classes. Finally, `BandstopFilter` shows marginally better results for a bandwidth of 1.2 Hz, although performance gains compared to the baseline are not significant.

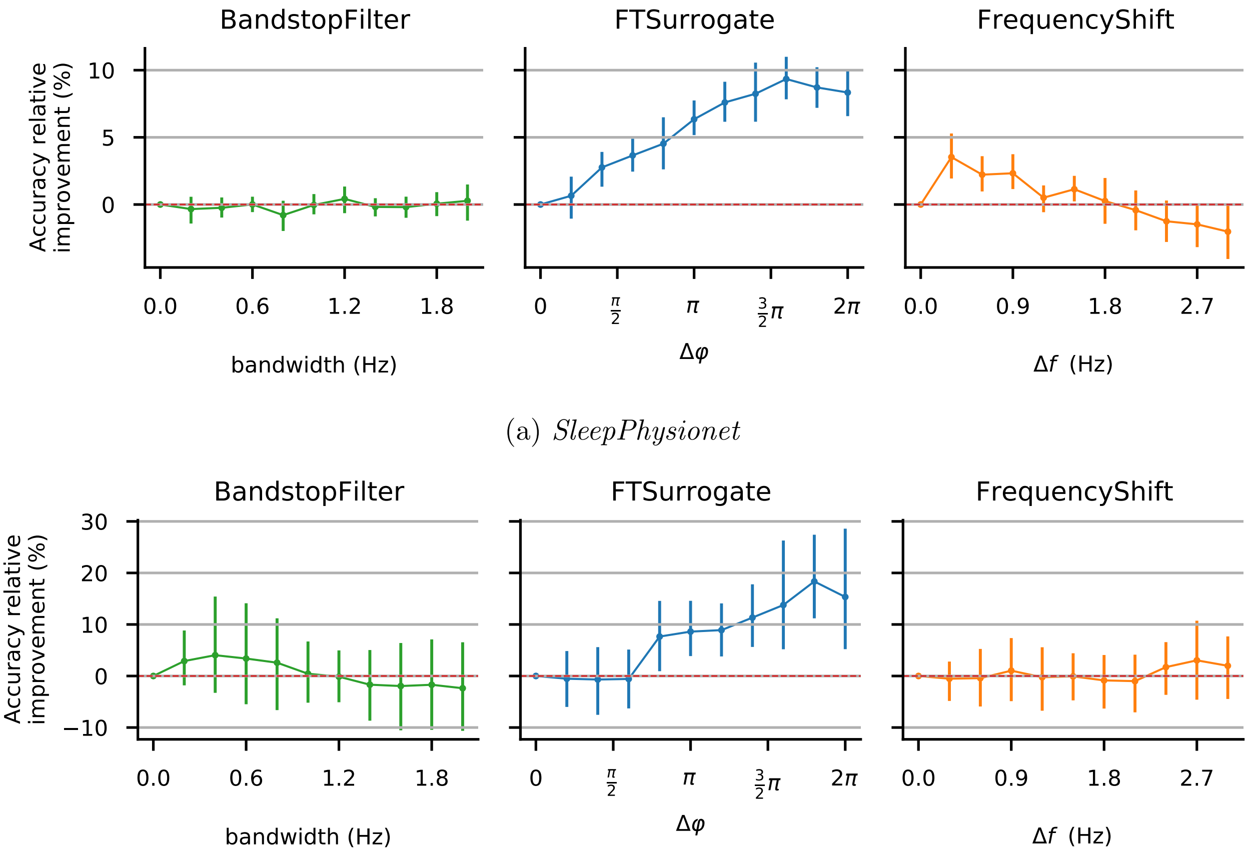
BCI IV 2a The results for *BCI IV 2a* presented in Figure 3b suggest several differences compared to *SleepPhysionet*. Unlike for the sleep stage classification task, smaller bandwidths seem to work better for `BandstopFilter`. Although, the high variance of the results suggest again that this transformation does not lead to statistically significant improvements. Likewise, maximum frequency shifts Δf privileged for this task are higher than for sleep staging. However, while `FrequencyShift` seems to significantly improve the predictive power with the *SleepPhysionet* dataset, the same cannot be said here. Since this transformation simulates inter-subject variability, we believe indeed that it is probably irrelevant for cross-session BCI, where models are trained and evaluated on the same subject. Finally, the `FTSurrogate` augmentation shows a pattern similar to the sleep staging task, as it benefits from higher $\Delta\varphi$ magnitudes.

3.2.2. Learning curves

SleepPhysionet As expected, the first learning curve experiment depicted in Figure 4a reveals that data augmentation methods are more helpful in low data regimes. It helps to mitigate the lack of data by artificially increasing the training set size. For example, a model trained with `FTSurrogate` on a fraction

Augmentation	Parameter	Interval	Unit	Best value (sleep staging)	Best value (BCI)
`BandstopFilter`	bandwidth	$[0, 2]$	Hz	1.2 Hz	0.4 Hz
`FTSurrogate`	$\Delta\varphi_{\max}$	$[0, 2\pi)$	rad	$\frac{4}{5}\pi$	$\frac{9}{10}\pi$
`FrequencyShift`	Δf	$[0, 3]$	Hz	0.3 Hz	2.7 Hz

Table 1: Adjustable parameter of each frequency domain augmentation.



(a) *SleepPhysionet*

(b) *BCI IV 2a*

Figure 3: Frequency augmentations parameters selection on the *SleepPhysionet* (a) and *BCI IV 2a* (b) datasets. Models were trained on respectively 350 and 60 windows using augmentations parametrized with 10 different linearly spaced values. Validation accuracies are reported relatively to a model trained without data augmentation. The error bars correspond to the 95% confidence intervals based on a 10-fold cross-validation.

smaller than 2^{-4} achieves performances comparable with a model trained without augmentation on twice as many data points. This first observation illustrates the interpretation of data augmentation as a regularization method, as it prevents the model from overfitting on a limited number of training examples. Moreover, the learning curves of models trained with `FrequencyShift` and `FTSurrogate` are above the baseline, thus evidencing that these augmentations preserve the relevant information of EEG signals for such a task, unlike

BandstopFilter, which appears to bring no improvements. Finally, a clear ranking stands out, revealing that the FTSurrogate augmentation yields the best improvements with a balanced accuracy relative gain of up to 10% in low data regimes.

BCI IV 2a Running the same experiment on the *BCI IV 2a* dataset yields different results as shown in Figure 4b. Unlike in sleep stage classification, BandstopFilter appears as a relevant data augmentation technique for this task, since it can improve accuracy by up to 17% in low data regimes. FrequencyShift also seems to be quite helpful for this dataset, even exceeding the improvements observed in sleep stage classification. This is surprising given that this transformation was originally designed to simulate the inter-subject variability observed in sleep stages, while the model is trained and evaluated on the same subject in cross-session BCI tasks. Figures 4a and 4b have in common the fact that FTSurrogate outperforms other frequency domain augmentations. For both classification tasks, a critical part of the information seems to lie in the frequency domain.

3.2.3. Per class analysis

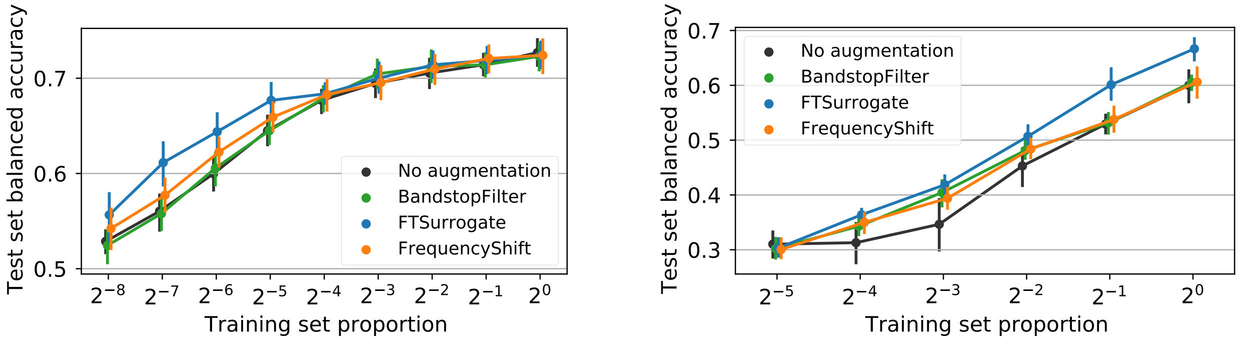
SleepPhysionet Taking a closer look at the performance per class presented in Figure 5a helps to reckon the variability of data augmentation effects with respect to the sleep stage. First, it can be seen that all frequency data augmentation methods only produce marginal improvements for sleep stages W and N3. This results can be interpreted in light of the performance of the baseline model presented in Figure 6a. Indeed, the highest F1-scores are reached for these classes and it is probably difficult to improve over the representations extracted by the baseline, which seem to already encode the relevant invariances. This experiments also discloses that frequency domain data augmentations significantly improve the results for the sleep stage REM. This result might not be only associated to the easiness of the task, since the baseline performs equally well on the N2 stage, which benefits less from data augmentation. It thus seems that the invariances encoded by such augmentation are specifically relevant for this class. Finally, the larger error bars for the REM stage may be associated with a higher inter-subject variability that might stem from the eye movement artifacts.

BCI IV 2a To the same extent, in Figure 5b the small improvements brought by frequency domain augmentations on the classification of right hand movements might be due to the already high performance reached by the baseline as shown in Figure 6b . It also appears that all three augmentations have almost the same effects on each class, aside from the variations related to the performance of the baseline. This fact points to similarities in frequencies that characterizes the brain activities for these 4 motor imagery tasks.

4. Time domain augmentations

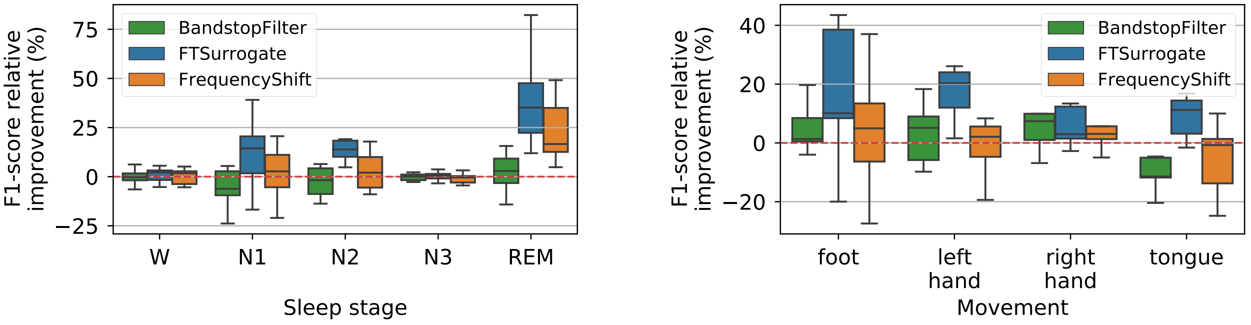
4.1. Rationale of the transformations

Gaussian noise EEG recordings are known to suffer from a limited signal to noise ratio. Besides noise does not affect equally all frequencies [34]. Low pass and high pass filtering are hence standard preprocessing steps, aiming to filter out some noise. While some research questions mostly focus on frequencies below 30 Hz, considering higher



(a) *SleepPhysionet* (b) *BCI IV 2a*

Figure 4: Learning curves for frequency domain augmentations along with the baseline trained with no augmentation. For each transformation, the same model is trained on 8 fractions of the dataset of increasing size. After each training, the average balanced accuracy score on the test set is reported with error bars representing the 95% confidence intervals estimated from 10-fold cross-validation.



(a) *SleepPhysionet* (b) *BCI IV 2a*

Figure 5: F1-score per class for frequency domain transformations. Scores are reported as relative improvement over a baseline trained without data augmentation. Models were trained on 180 and 230 time windows for *SleepPhysionet* and *BCI IV 2a* datasets respectively. Boxplots were estimated from 10-fold cross-validation.

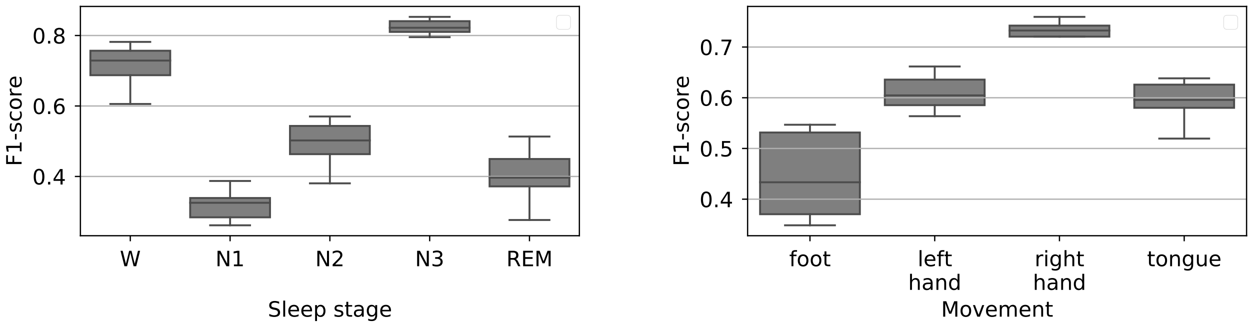
frequencies as noise, some others investigate high frequency oscillations ($\geq 80\,\mathrm{Hz}$) [35].

The inability to get rid of the noise inherently present in brain signals motivates the introduction of the `GaussianNoise` augmentation, which mimics this feature and promotes robustness to EEG acquisition noise [36]. This transformation consists in adding Gaussian white noise $E(t) \sim \mathcal{N}(0, \sigma)$ to the original signal X,

$$\texttt{GaussianNoise}[X](t) = X(t) + E(t),$$

where σ denotes the standard-deviation of the Gaussian distribution.

As EEG signal power decreases with the frequency, it mostly preserves power band ratios at lower frequencies, which are instrumental in many EEG decoding tasks, such as sleep stage classification [37, 5].



(a) *SleepPhysionet*

(b) *BCI IV 2a*

Figure 6: F1-score per class for a baseline model trained without data augmentation. Boxplots were estimated from 10-fold cross-validation.

However, by adding the same amount of power to all frequencies, the addition of white noise hides the information contained in high frequency bands, which have smaller power, as depicted in Figure 7. From this perspective, the effect of this transformation is somehow analogue to a low-pass filter where the parameter σ stands for a cut-off frequency (*cf.* Figure 7).

Smooth time mask As described in the *AASM* scoring manual [21], sleep stages are most often characterized by the global information contained in a time window. For example, the sleep stage N1 is scored when more than 15 seconds ($\geq$ 50%) of a time window is dominated by theta activity ($4 - 7\,\mathrm{Hz}$). The representations learned by the model should thus encapsulate the global information of the signal and avoid to rely on transient patterns. Consequently, one would expect two almost identical EEG windows differing only for a few seconds to be very close in the representation space [32]. The `SmoothTimeMask` augmentation is designed with this in mind [18]. It consists in replacing by zeroes a portion of length Δt, that starts from a randomly sampled instant t_{cut}. The computation is carried out by multiplying the signal X by a mask m_λ, which is made out of two opposing sigmoid functions of temperature λ:

$$\begin{aligned}
&\texttt{SmoothTimeMask}[X](t) := X(t) \cdot m_\lambda(t), \\
&m_\lambda(t) := \sigma_\lambda(t - t_{\text{cut}}) + \sigma_\lambda(t_{\text{cut}} + \Delta t - t) \\
&\sigma_\lambda(t) := \frac{1}{1 + \exp\left(-\lambda t\right)} \ , \\
&t_{\text{cut}} \sim \mathcal{U}[t_{\min}, t_{\max} - \Delta t] \ .
\end{aligned}$$

This allows to set the signal smoothly to zero, as shown in Figure 8, and avoids creating discontinuities.

By masking part of the signal, we assign the same label (*e.g.*, sleep stage or action) to windows differing in the time domain only inside a short time span. This transformation hence promotes such an invariance, and intends to teach the feature extractor to be more robust to differences at this time scale.

Sign flip The electric potentials measured with EEG are driven by post-synaptic potentials along dendrites of pyramidal neurons. Depending on the geometric alignment of active neurons, the current produced can add up to be measured non-invasively with EEG. The group of neurons can be well modeled as elec-

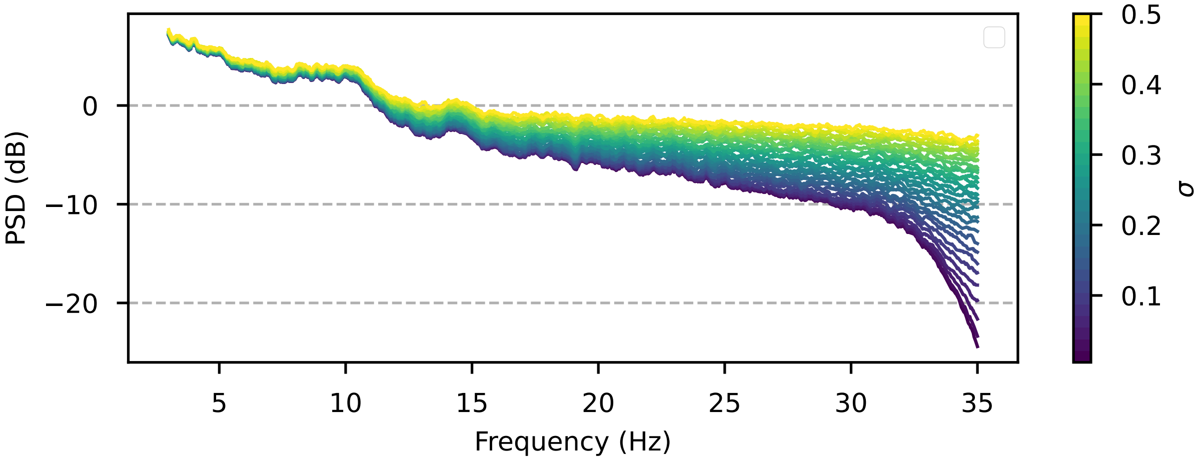


Figure 7: Effects of the addition of `GaussianNoise` on the PSD. Power spectra were averaged over N1 windows for one night of sleep from the *SleepPhysionet* dataset. In polysomnographies, the power globally decreases as the frequency increases. Consequently, the `GaussianNoise`, which adds a constant amount of power across all frequencies, has a greater relative impact on higher frequencies. As we increase σ, a greater portion of the signal is hidden by the added noise.

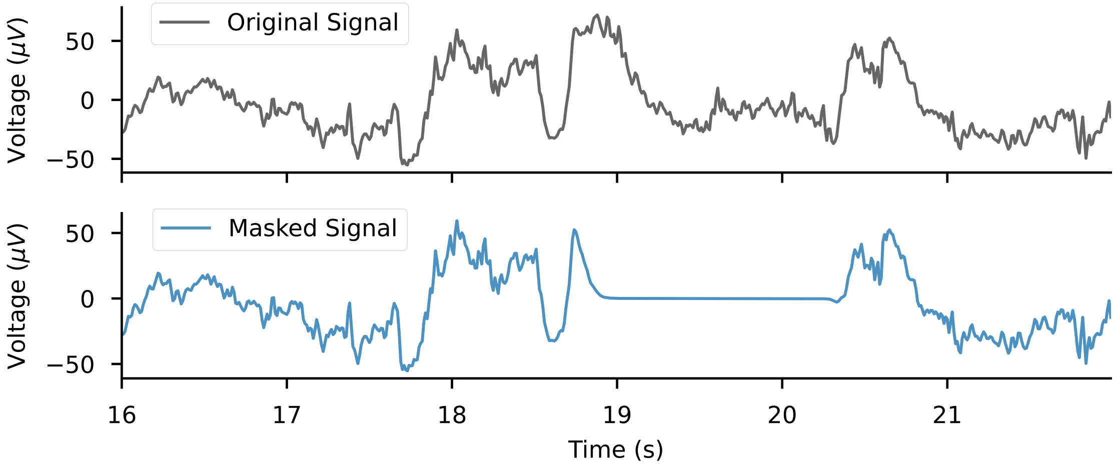


Figure 8: Effect of `SmoothTimeMask` on a time window from the *SleepPhysionet* dataset. The mask length is $1.6 sec$ and the transition between unchanged parts and the masked portion is smooth.

tric current dipole, characterized by a moment vector at a given location in the brain. For most analysis, the moment strength (norm) and location (origin) are sufficient. While the moment direction is often unused, it is responsible for the sign of the potential measured by the EEG device. Our guess is that changing the sign of EEG channels will preserve the instrumental information contained in the strength and location of the dipole. In terms of physiology, it corresponds to current flowing from superficial cortical layers to deep layers or vice versa. Moreover, the choice of the reference electrode can also affect the polarity of the EEG potentials.

This motivated the introduction of the `SignFlip` augmentation [29], which simply multiplies each EEG channel by -1 with probability p_{aug} (*cf.* Section 2.1):

$$\texttt{SignFlip}[X](t) := -X(t).$$

This transformation preserves the topographical properties of the electric field potential since it corresponds to a swap between the head and tail of the dipole, without changing its location.

Time reverse Sleep stages are mainly characterized by frequency-domain information [21]. Considering that the orientation of the time axis has no effect on the signal's PSD, it can be hypothesized that flipping the time axis preserves most of the information while generating a new input. With this in mind, the `TimeReverse` augmentation was proposed in [29] and is simply implemented as follows:

$$\texttt{TimeReverse}[X](t) := X(t_{\text{max}} - t),$$

where t_{max} is the length of a time window.

By promoting invariance to the direction of the time axis, not only does this augmentation leave the frequency domain unchanged, it also preserves the proportions of certain rhythms within the signal. This is probably a desirable property in sleep stage classification tasks, where some stages, such as N1, are scored based on the dominant rhythms observed (theta waves).

Moreover, a large part of the time-domain information is preserved by this transformation, since symmetric waveforms are merely shifted along the time axis and only asymmetric patterns are modified, as illustrated in Figure 9. Note also that this transformation also implies partial invariance to the position of transient temporal patterns within the EEG signals. This might be a useful property for instance to score the sleep stage N2, which is characterized by more than two occurrences of K-complexes throughout the stage [21].

4.2. Experimental results

4.2.1. Parameters selection As in Section 3.2.1, here we investigate the effects on the performance of the strength of previously introduced transformations. As shown in Table 2, the intensity of `GaussianNoise` is controlled by its standard deviation σ, while `SmoothTimeMask` is governed by the length of the mask Δt. Note that there is no notion of strength or intensity for `TimeReverse` and `SignFlip`, which are hence not studied in this subsection.

SleepPhysionet Similarly to what can be observed for frequency transformations, the impact of the intensity is quite different between `SmoothTimeMask` and `GaussianNoise`, as illustrated in Figure 10. While increasing the strength of `SmoothTimeMask` seems beneficial, no clear trend is observed for `GaussianNoise`. For `SmoothTimeMask`, we restricted our experiment to masks of less than two seconds to avoid removing too much crucial information from the signal. It seems that the augmentation is more efficient with masks of maximum length.

BCI IV 2a On the one hand, the grid search on the *BCI IV 2a* dataset presented in Figure 10b shares similarities with its counterpart

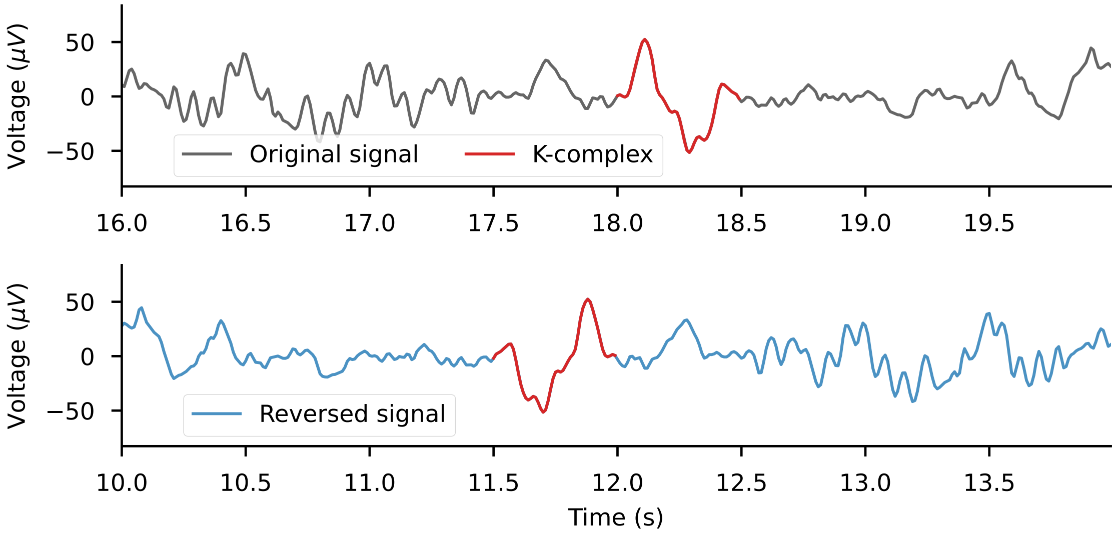


Figure 9: `TimeReverse` augmentation on an EEG signal from the *SleepPhysionet* dataset. A large part of the signal is not deeply affected. Wave patterns are merely translated along the time axis. Some specific asymmetric EEG patterns such as K-complexes are reversed by this transformation.

on the *SleepPhysionet* dataset (Figure 10a). In both cases, increasing the mask length enhances the performance of `SmoothTimeMask`, whereas `GaussianNoise` does not seem to lead to any robust improvement. On the other hand, `SmoothTimeMask` appears to be much more useful on the BCI task since it produces relative improvements of up to 20% when only 60 windows per class are used for the training.

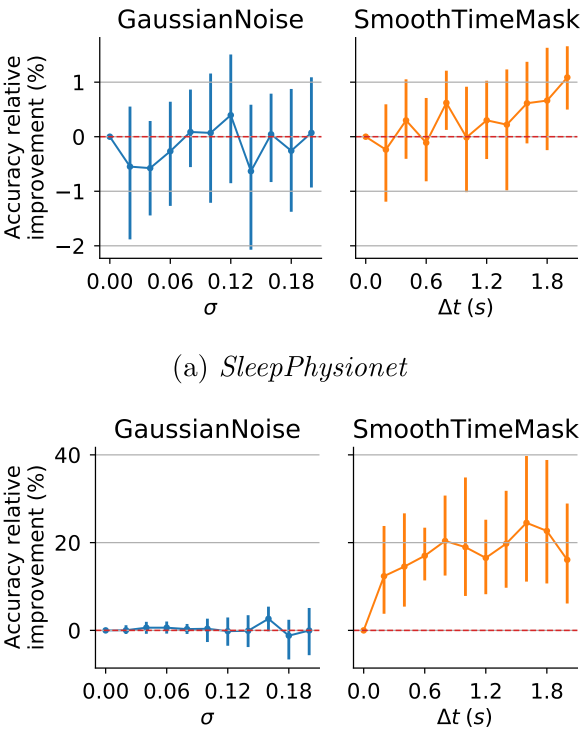
4.2.2. Learning curves

SleepPhysionet The learning curves for the sleep staging task presented in Figure 11a reveals that the most significant improvements are brought by `SignFlip` and `TimeReverse` with the latter outperforming the former. As expected from the results of the previous experiment, `GaussianNoise` and `SmoothTimeMask` have a negligible effect on this sleep staging task and perform on par with the baseline model. These results suggest that symmetries are notably relevant invariances for the sleep stage classification task. They preserve both the frequencies and some of the transient patterns occurring during polysomnographies such as sleep spindles which presents a symmetry both along the x and y axis. This observation also supports the claim that sleep scoring heavily relies on frequency domain features since the two augmentations that leave the power spectrum density of the signal unchanged outperform the others.

BCI IV 2a Figure 11b contains the learning curve plots for the *BCI IV 2a* dataset. It suggests that `TimeReverse` and `SmoothTimeMask` are the most suited time domain augmentations for the BCI task in low and high data regimes respectively. An intuitive interpretation of this result is that in motor imagery, subjects are asked to mentally simulate the same physical action during the whole trial

Augmentation	Parameter	Interval	Unit	Best value (sleep staging)	Best value (BCI)
`GaussianNoise`	σ	$[0, 0.2]$	-	0.12	0.16
`SmoothTimeMask`	Δt	$[0, 2]$	s	2 s	1.6 s

Table 2: Adjustable parameter for each time domain augmentation.



(a) *SleepPhysionet*

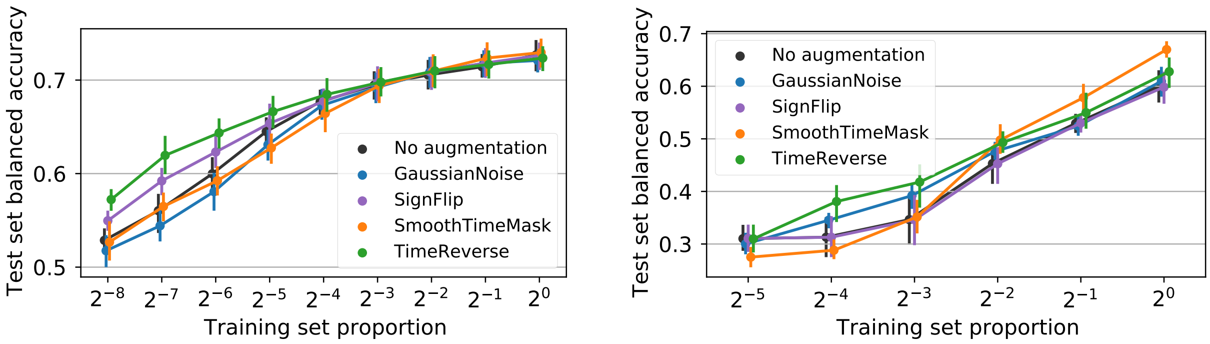
(b) *BCI IV 2a*

Figure 10: Time augmentations parameters selection on the *SleepPhysionet* (a) and *BCI IV 2a* (b) datasets. Models were trained on respectively 350 and 60 windows using augmentations parametrized with 10 different linearly spaced values. Validation accuracies are reported relatively to a model trained without data augmentation. The error bars correspond to the 95% confidence intervals based on a 10-fold cross-validation.

and the information encoded in brain signals should hence be invariant to local distortions. Another striking observation has to do with the learning curve of `SignFlip`, which brings no improvement over the baseline while it is the second most efficient augmentation for the sleep staging task. In the BCI competition, EEGs were recorded using a 22 electrodes montage, with the left mastoid as reference for all of them [25], while the two electrodes used in *SleepPhysionet* have different references. This may be a possible reason for the divergent results given the interpretation of `SignFlip` as switching the reference electrodes (*cf.* section 4).

4.2.3. Per class analysis

SleepPhysionet Figure 12a introduces the results per class of the time domain augmentations for the sleep staging task. As with frequency domain augmentations, the REM stage seems to benefit more than the others from this class of transformations. This stage is characterized by low amplitude mixed frequency and bursts of eye activity. The scoring thus relies heavily on global amplitudes which are mostly preserved with time domain augmentations (except for `GaussianNoise` which performs the worst among them). `TimeReverse` especially yields the best results for this class as well as for all others. Note also that while `SmoothTimeMask` leads to improvements for REM and W stages, it appears to be detrimental for learning to recognize non-REM sleep. A possible interpretation of this observation is that this transformation may erase important waves, such as K-complexes and spindles, which strongly characterize stages N2 and N3.



(a) *SleepPhysionet*

(b) *BCI IV 2a*

Figure 11: Learning curves for time domain augmentations along with the baseline trained with no augmentation. For each transformation, the same model is trained on fractions of the dataset of increasing size. After each training, the average balanced accuracy score on the test set is reported with error bars representing the 95% confidence intervals estimated from 10-fold cross-validation.

BCI IV 2a Figure 12b introduces the results per class of the time domain augmentations for the BCI motor imagery task. While `SignFlip` and `GaussianNoise` do not seem to help for any imagined movement, `TimeReverse` and `SmoothTimeMask` greatly improve the predictive accuracy for most classes, specially left-hand, right-hand and foot. Namely, `SmoothTimeMask` outperforms all frequency domain augmentations for the classification of all classes except tongue movements.

5. Spatial domain augmentations

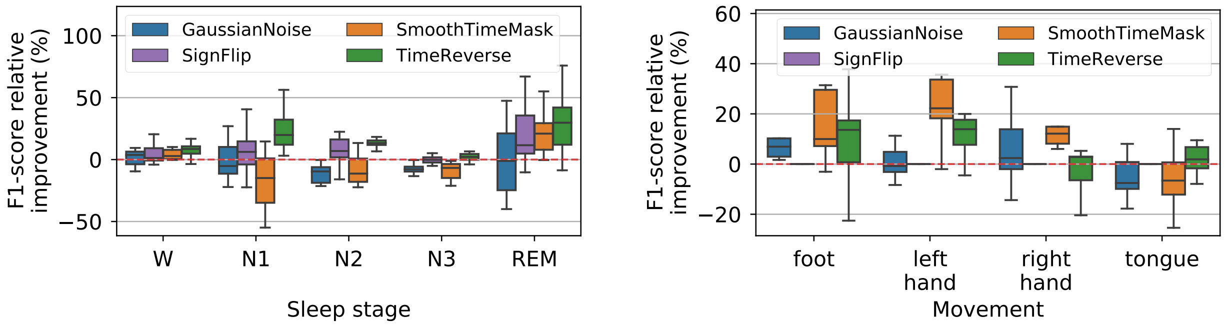
In this section we study augmentations exploiting the sensors spatial positions.

5.1. Rationale of the transformations

Channels symmetry Several brain activities that are monitored using EEG involve a sagittal plane (left-right) symmetry. For example, it has been evidenced that tongue movements stem from an activation of the primary and supplementary sensorimotor areas without any significant lateralization [38]. Taking this fact into consideration, the `ChannelsSymmetry` augmentation, proposed in [39], permutes the order of EEG channels in order to simulate a swap between EEG sensors placed on the right and left hemispheres.

This augmentation hence teaches the model to become invariant to the position of sensors with regard to the sagittal plane, hence helping to extract representations which rely on brain activity that is not lateralized. Yet, some neural activities are strongly lateralized, for example with BCI tasks. Indeed, since the seminal work of Penfield and Jasper in 1954 [40], it is well known that the primary sensory responses are contralateral, meaning for example that a touch with the left hand results in strong neural activations on the right hemisphere. Consequently, applying the `ChannelsSymmetry` transformation is expected to be harmful to recognize categories such as left vs. right hand movements.

Channels dropout A major hurdle for the analysis of EEG signals is the inconsistent quality of EEG channels throughout a record-



(a) *SleepPhysionet* (b) *BCI IV 2a*

Figure 12: F1-score per class for time domain transformations. Scores are reported as relative improvement over a baseline trained without data augmentation. Models were trained on 180 and 230 time windows for *SleepPhysionet* and *BCI IV 2a* datasets respectively. Boxplots were estimated from 10-fold cross-validation.

ing. For example, in polysomnography, changes in the subject's position during sleep might result in a loss of contact between several electrodes and the scalp. Beyond sleep applications, the spread of mobile wearable EEG devices raises new challenges, as they are more prone to noise and missing channels [41]. Finally, the EEG and machine learning communities consider with great interest the question of transferability across datasets, which raises major challenges regarding inconsistent numbers of channels or channels ordering.

To tackle these issues, the `ChannelsDropout` augmentation, initially proposed in [19], randomly sets each channel of the EEG recording to zero with a given probability p_{drop}. More precisely, for $X \in \mathbb{R}^{C \times T}$ an EEG window of T samples collected on C channels, the augmented signal takes the form

$$\texttt{ChannelsDropout}[X]_c := d_c \cdot X_c,$$

where d_c's are sampled from a Bernoulli distribution $\mathcal{B}$ of probability p_{drop}, and X_c denotes the c^{th} row of X corresponding to channel $c \in \{1, \dots, C\}$.

As the widely used dropout layer [16], this augmentation prevents the model from relying too heavily on a given input channel, which could lead to overfitting, poor generalization on different datasets and lack of robustness to corrupted channels.

More precisely, by reproducing defective input channels which naturally arise in EEG recordings, this augmentation method aims at teaching models to become invariant to the number of channels. Incorporating this invariance should ensure that the model learns from the global information available from all channels instead of relying on a single one.

Channels shuffle Another augmentation called `ChannelsShuffle` is also proposed in [19] with the same objective of making EEG classification models more robust to cross-dataset transfer. It consists in randomly permuting the rows of EEG input matrices

$$\texttt{ChannelsShuffle}[X]_c := X_{\tau(c)}, \tag{1}$$

where τ is uniformly sampled from all the possible permutations. Although all channels are shuffled in the original formulation proposed in [19] ($I = \{1, \dots, C\}$), we implement this augmentation with a variable subset of permuted channels, controlled using a settable probability p_{shuffle} of adding each channels to the permutation set I. Our formulation hence amounts to the same transformation in Equation 1, where permutations τ are sampled from a random subset $I = \{c | s_c = 1, s_c \sim \mathcal{B}(p_{\text{shuffle}})\}$, where $\mathcal{B}(p)$ denotes a Bernoulli distribution with probability p. This new parameter p_{shuffle} hence allows to define how strongly the input signals will be transformed, while reproducing the original formulation when $p_{\text{shuffle}} = 1$.

In addition to helping the transfer to datasets with different channels ordering, this augmentation induces invariance to the absolute and relative positioning of EEG sensors, since it prevents the decision function from relying on it. Such a transformation can make sense for several EEG classification tasks for which the precise localization of the cerebral activity is not strongly predictive, such as sleep staging. Indeed, sleep experts hardly consider the sensors position of the channel they observe (*e.g.,* Fpz-Oz) but rather rely on the waveforms and spectral characteristics (*e.g.,* theta activity, K-complex).

If this augmentation seems well suited for the aforementioned task, it is expected to be less efficient in a context where sensors positions and source localization features have a greater impact, such as BCI.

Sensors rotations Across the acquisition of EEG data, *e.g.,* between different recording sessions, the cap can move over the subjects head, resulting in small perturbations on sensors locations.

In an attempt to make EEG classifiers more robust, `SensorsRotation` augmentation simulates the natural position variations expected during the acquisition [20]. To achieve this, the electrical potentials measured during an EEG recording are interpolated in between sensors using their 3D coordinates in a standard 10-20 montage. The interpolated signals hence approximate what would have been recorded with a device slightly rotated along a given axis (x, y or z). While in [20] a radial-basis functions interpolator is used, we decided to implement this transformation using spherical splines, as commonly done for bad EEG channels interpolation [42]. This was implemented using the MNE-Python library [33].

Compared to `ChannelsShuffle`, which encourages *global* invariance to electrodes positions, `SensorsRotation` induces robustness to small and local variations of sensors' positions.

5.2. Experimental results

5.2.1. Parameters selection The parameters listed in Table 3 control the strength of the transformations, namely: the probability to drop channels p_{drop}, the probability to shuffle channels p_{shuffle} and the angle of rotation θ_{rot} respectively for `ChannelsDropout`, `ChannelsShuffle` and `SensorsRotations`. Regarding the sensors rotations, we restricted the range of possible angles to $[0, 30]$ degrees as done in [20]. The strength of the `ChannelsSymmetry` augmentation cannot be adjusted.

SleepPhysionet The results of the grid search for the parameters of spatial augmentations are presented in Figures 13a and 14a. We can see that `SensorsRotations` consistently lead to poor performances. This can be explained by the scarce number of EEG sensors available in this dataset, resulting in imprecise interpolation. This same reason might also explain why high probabilities of dropping channels in `ChannelsDropout` appear to be

Augmentation	Parameter	Interval	Unit	Best value (sleep staging)	Best value (BCI)
`ChannelsDropout`	p_{drop}	$[0, 1]$	-	0.4	1
`ChannelsShuffle`	p_{shuffle}	$[0, 1]$	-	0.8	0.1
`SensorXRotations`	θ_{rot}	$[0, 30]$	degree	25^o	3^o
`SensorYRotations`	θ_{rot}	$[0, 30]$	degree	9^o	12^o
`SensorZRotations`	θ_{rot}	$[0, 30]$	degree	30^o	3^o

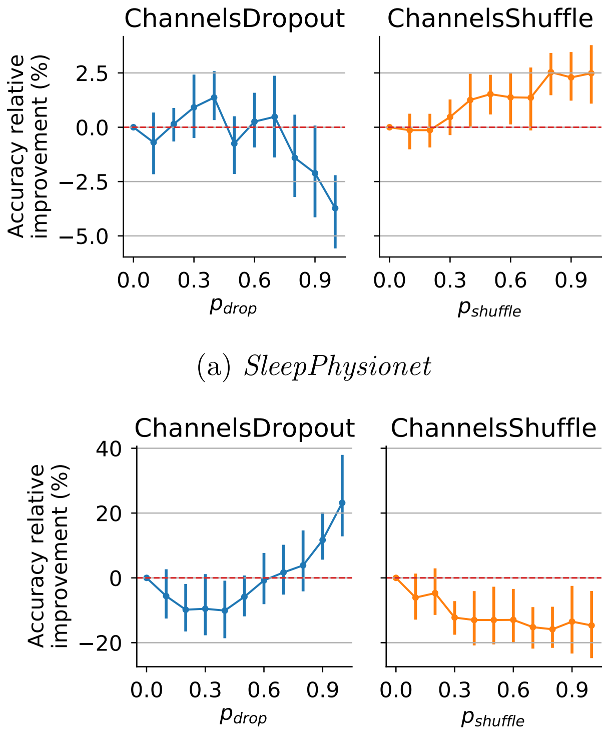
Table 3: Potential and selected values for the adjustable parameter of each spatial domain augmentation.

detrimental to learning, the best results being obtained with $p_{\text{drop}} = 0.4$. Indeed, as there are only two channels in this dataset, a probability of $p_{\text{drop}} = 0.7$ would erase all channels for 1 out of 4 windows on average (the probability to augment $p_{\text{aug}} = 0.5$, multiplied by p_{drop} squared). On the contrary, for `ChannelsShuffle`, higher probability to shuffle channels yields the best results. This was to be expected, since sleep stage information is not very spatially localized.

BCI IV 2a As shown in Figure 13b, the strength of the spatial domain augmentations impacts the performance on the motor imagery task in a completely different way than on the sleep staging task. Indeed, the patterns for `ChannelsDropout` and `ChannelsShuffle` are reversed here: larger values of p_{drop} yield better performances, while `ChannelsShuffle` is consistently harmful, with stronger shuffling probabilities yielding the worst performances. The order of magnitude of the impact of these augmentations is also different compared to the sleep staging case, with up to 20% improvement for `ChannelsDropout`. Moreover, it is surprising to obtain the best results with a probability $p_{\text{drop}} = 1$, given that it corresponds to all channels being dropped for 1 out of 2 windows. This might indicate that our model is overfitting the data, since gradients computed from fully dropped examples only update the biases of the model. Regarding the rotational augmentations, parameter selection results are depicted in Figure 14b and are hardly readable because of the very large error bars and mean values close to zero. It however seems that larger rotational angles around the Y and Z axis tend to have a slightly negative effect on the performance, while rotations around the X axis have small positive effects on average.

5.2.2. Learning curves

SleepPhysionet The sleep staging learning curves of spatial domain augmentations are plotted on Figure 15a. All augmentations seem to globally perform on par with the baseline. It thus seems that such augmentations are not particularly relevant for the sleep staging task, at least on the *SleepPhysionet* dataset. This was expected since this dataset only contains 2 EEG channels and hence, does not encode much spatial information. Note that the `ChannelsSymmetry` augmentation was omitted for this experiments since it would correspond to the Identity mapping, as both EEG electrodes are on the sagital plane here.



(a) *SleepPhysionet*

(b) *BCI IV 2a*

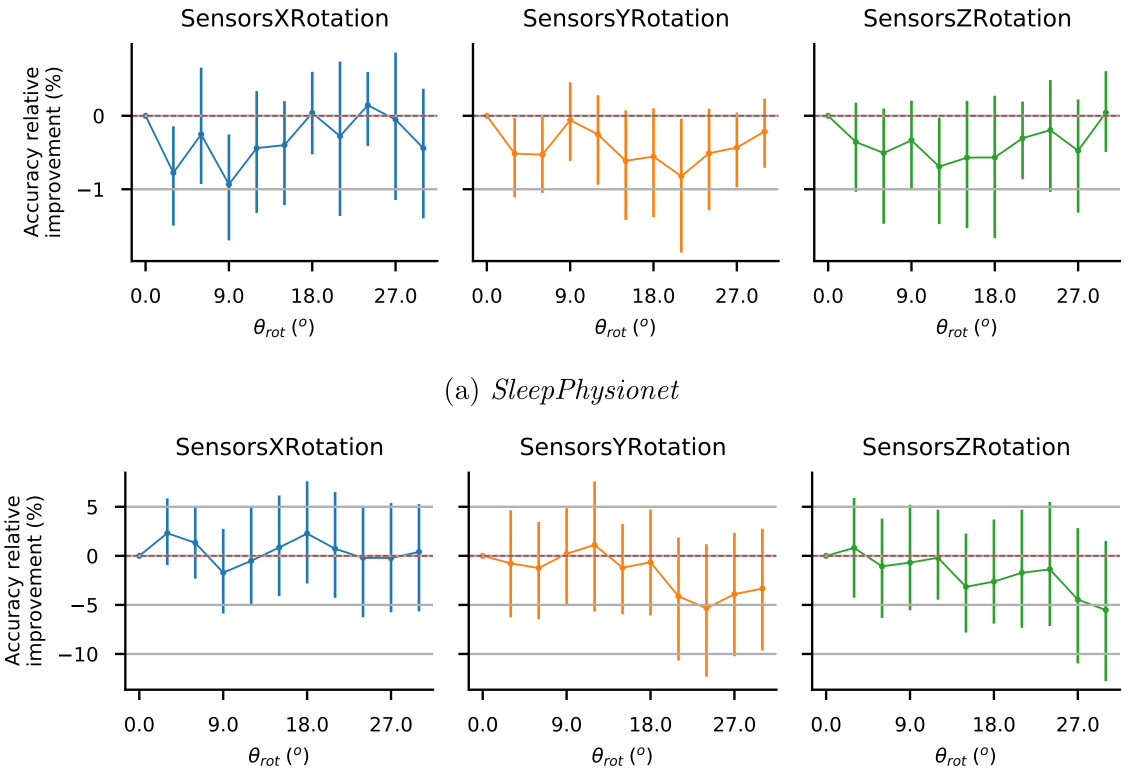
Figure 13: Spatial augmentations parameters selection on the *SleepPhysionet* (a) and *BCI IV 2a* (b) datasets. Models were trained on respectively 350 and 60 windows using augmentations parametrized with 10 different linearly spaced values. Validation accuracies are reported relatively to a model trained without data augmentation. The error bars correspond to the 95% confidence intervals based on a 10-fold cross-validation.

BCI IV 2a The learning curves obtained with spatial data augmentations are presented on Figure 15b. They confirm that mixing channels up with `ChannelsShuffle` and `ChannelsSymmetry` do not help in low-data regimes and might be detrimental for the motor imagery task when there is enough data for learning. This observation can be related to the well known crucial importance of the spatial information for motor imagery tasks. On the contrary, `ChannelsDropout` augmentation significantly enhance the performance, specially when there is enough data to keep learning despite dropping part of the information. Unlike the two previous spatial augmentations, `ChannelsDropout` might help to learn more robust motor imagery features by hiding part of the spatial information without misleading the model. Finally, concerning the `SensorsRotation` augmentations, the results depicted in Figure 16 confirm that sensors rotations hardly have any statistically significant effect on the performance. These results were expected after the parameter search experiment (Figure 14b) which showed that the optimal angles of rotation yield almost no improvement on average.

5.2.3. Per class analysis

SleepPhysionet While Figure 17a confirms that spatial augmentations have no significant impact on the classification of stages W, N1 and N2, it also brings more nuance to the previous results from Figure 15a. Indeed, it seems that even with low probabilities, dropping channels can significantly harm the recognition of N3 and REM stages, which might indicate that those are more easily identified on one of the two available channels. Likewise, shuffling seems to degrade the performance for the N3 stage, while yielding 10% median improvement for REM stages. This might indicate that N3 stages are partly characterized by spatial patterns which are lost when channels are shuffled. It also seems to confirm that REM stages are not localized and rather correspond to a global brain activity, similar to the awake state.

BCI IV 2a The results per class for the motor imagery task presented in Figure 17b confirm



(a) *SleepPhysionet*

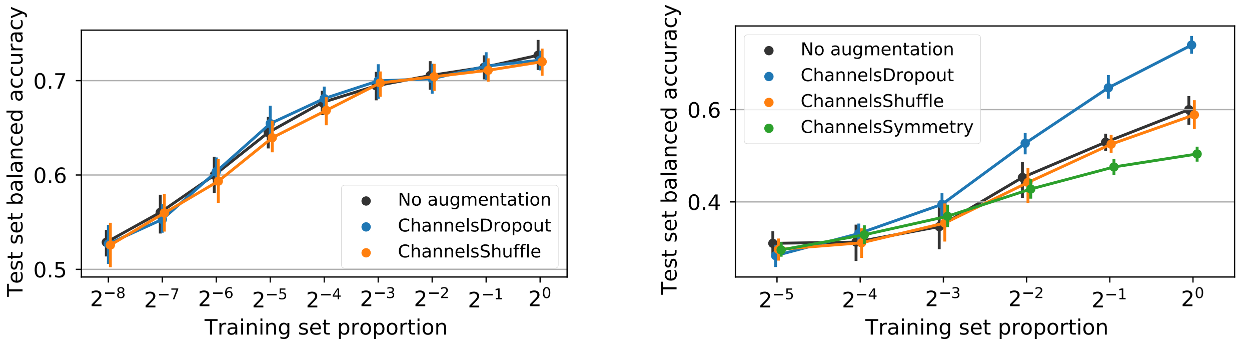
(b) *BCI IV 2a*

Figure 14: Rotational augmentations parameters selection on the *SleepPhysionet* (a) and *BCI IV 2a* (b) datasets. Models were trained on respectively 350 and 60 windows using augmentations parametrized with 10 different linearly spaced values. Validation accuracies are reported relatively to a model trained without data augmentation. The error bars correspond to the 95% confidence intervals based on a 10-fold cross-validation.

that `ChannelsSymmetry` is particularly detrimental for right and left hand movements, which are characterised by a heavily lateralised brain activity. Meanwhile, this augmentation slightly helps to learn to recognize tongue movements, which are associated with non-lateralised brain activities. In fact, this last class seems to be the least affected by all three augmentations, which can be explained by the weaker importance of spatial information for this class [38]. Concerning spatial rotations, Figure 18 also helps to nuance the previous aggregated results from Figure 16. Indeed, we see that although boxes' whiskers reach negative values, rotations around the Z axis consistently improve the performance in at least 75% of cases by a significant amount.

6. Conclusion

By allowing to increase the training data during learning, data augmentation limits the need for large annotated datasets that are required to fully leverage the potential of deep learning models. In this paper we carried out a unified and quasi-exhaustive analysis of existing data augmentation methods for EEG signals. To this end, we have presented



(a) *SleepPhysionet*

(b) *BCI IV 2a*

Figure 15: Learning curves for spatial domain augmentations (except rotations) along with the baseline trained with no augmentation. For each transformation, the same model is trained on fractions of the dataset of increasing size. After each training, the average balanced accuracy score on the test set is reported with error bars representing the 95% confidence intervals estimated from 10-fold cross-validation.

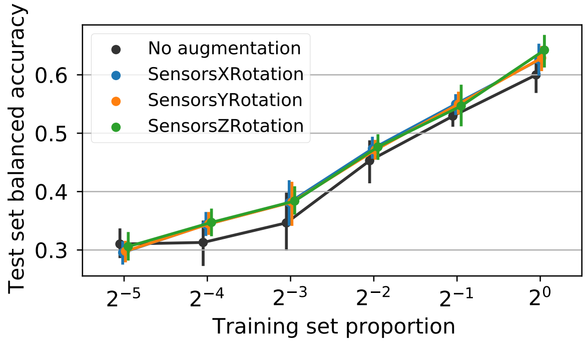
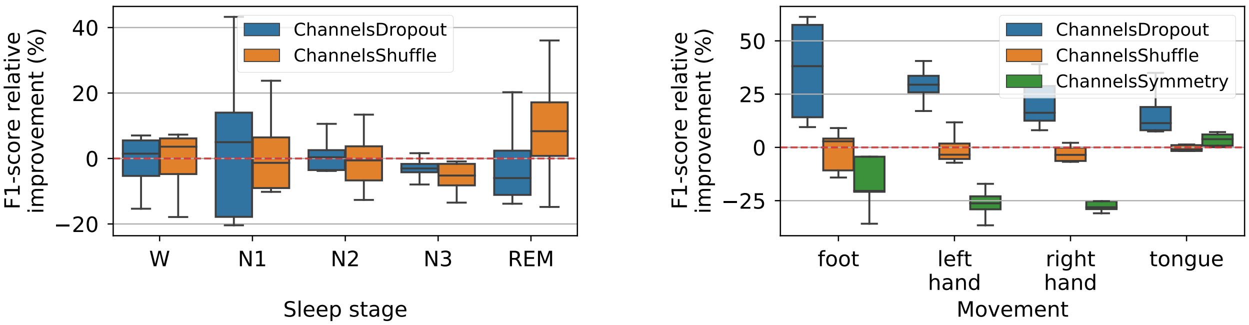


Figure 16: Learning curves for sensors rotations augmentations on the *BCI IV 2a* dataset, along with the baseline trained with no augmentation. For each transformation, the same model is trained on fractions of the dataset of increasing size. After each training, the average balanced accuracy score on the test set is reported with error bars representing the 95% confidence intervals estimated from 10-fold cross-validation.

the rationale and assumptions behind each augmentation considered, both from the perspective of the underlying neurophysiology and of the experimental setups. Overall, our experimental results demonstrate that the use of data augmentation is beneficial for the training of EEG classifiers, both for sleep staging and BCI tasks, yielding sometimes more than 10% improvements, specially in low-data regimes. Importantly, illustrating these results on two substantially different datasets highlights the diversity of the relevant data augmentations. In particular, our experiments demonstrate the importance of the selection of both the right transformation and strength for each different type of task considered. While time-frequency transformations appear to be preferable for sleep stage classification, spatial augmentations also seem competitive for motor imagery tasks. Moreover, our per-class analysis allowed to identify structural differences between different imagined actions and sleep stages, thus also demonstrating the descriptive usefulness of augmentations. While this study is not completely exhaustive and could be enriched by the addition of other datasets and new augmentations, we believe that the methodological framework presented, along with our reproducible code,



(a) *SleepPhysionet* (b) *BCI IV 2a*

Figure 17: F1-score per class for spatial domain transformations (except for rotations). Scores are reported as relative improvement over a baseline trained without data augmentation. Models were trained on 180 and 230 time windows for *SleepPhysionet* and *BCI IV 2a* datasets respectively. Boxplots were estimated from 10-fold cross-validation.

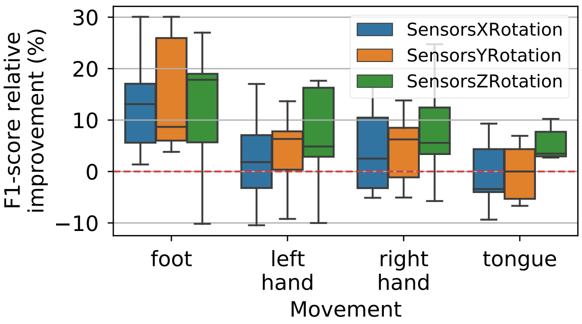


Figure 18: F1-score per class for rotation transformations. Scores are reported as relative improvement over a baseline trained without data augmentation. Models were trained on 230 time windows of the *BCI IV 2a* dataset. Boxplots were estimated from 10-fold cross-validation.

should allow to conduct similar analysis on any other EEG dataset and augmentation. We believe that this systematic analysis of EEG data augmentation will help practitioners improve their predictive models and foster new research works aiming to better understand augmentation methods for EEG signals.

References

[1] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5):051001, 2019.

[2] Junjian Chen, Zhuliang Yu, Zhenghui Gu, and Yuanqing Li. Deep temporal-spatial feature learning for motor imagery-based brain–computer interfaces. 28(11):2356–2366, 2020.

[3] Huy Phan, Oliver Y. Chen, Minh C. Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[4] Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-sleep: resilient high-frequency sleep staging. *NPJ digital medicine*, 4(1):1–12, 2021.

[5] Stanislas Chambon, Mathieu Galtier, Pierrick Arnal, Gilles Wainrib, and

Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769, 2018.

[6] Kai Keng Ang, Zheng Yang Chin, Chuanchu Wang, Cuntai Guan, and Haihong Zhang. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers in neuroscience*, 6:39, 2012.

[7] Lukas A. W. Gemein, Robin T. Schirrmeister, Patryk Chrabaszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of EEG pathology. *NeuroImage*, 220:117021, 2020.

[8] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. 8(4):441–446, 2000.

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

[10] Richard S. Rosenberg and Steven Van Hout. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, 9(1):81–87, 2013.

[11] Maureen Clerc, Laurent Bougrain, and Fabien Lotte. *Brain-Computer Interfaces 1*. Wiley-ISTE, July 2016.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NeurIPS)*, 2012.

[13] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA, 2019.

[14] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for NLP. *arXiv:2105.03075*, 2021.

[15] Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A Group-Theoretic Framework for Data Augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. 15(56):1929–1958, 2014.

[17] Justus T. C. Schwabedal, John C. Snyder, Ayse Cakmak, Shamim Nemati, and Gari D. Clifford. Addressing Class Imbalance in Classification Problems of Noisy Signals by using Fourier Transform Surrogates. *arXiv:1806.08675*, 2019.

[18] Mostafa Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive Representation Learning for Electroencephalogram Classification. In *Machine Learning for Health*, 2020.

[19] Aaqib Saeed, David Grangier, Olivier Pietquin, and Neil Zeghidour. Learning from heterogeneous EEG signals with differentiable channel reordering. In

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[20] Mario Michael Krell and Su Kyoung Kim. Rotational data augmentation for electroencephalographic data. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 471–474. IEEE, 2017.

[21] R.B. Berry, R. Brooks, C.E. Gamaldo, S.M. Harding, R.M. Lloyd, C.L. Marcus, B.V. Vaughn, and American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications : Version 2.3.*

[22] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. Publisher: Am Heart Assoc.

[23] Allan Rechtschaffen and Anthony Kales. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Brain Information Service/Brain Research Institute, University of California.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[25] Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, 16:1–6, 2008.

[26] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

[27] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 2017.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR).*

[29] Cédric Rommel, Thomas Moreau, Joseph Paillard, and Alexandre Gramfort. CADDA: Class-wise Automatic Differentiable Data Augmentation for EEG Signals. *arXiv preprint arXiv:2106.13695*, 2021.

[30] R. T. Canolty et al. High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, 313(5793):1626–8, 2006.

[31] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.

[32] Joseph Y. Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-Aware Contrastive Learning for Biosignals. *arXiv:2007.04871*, 2020.

[33] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG Data Analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013.

[34] Mike X Cohen. *Analyzing neural time series data: theory and practice.* MIT press, 2014.

[35] Birgit Frauscher, Fabrice Bartolomei, Katsuhiro Kobayashi, Jan Cimbalnik, Maryse A van 't Klooster, Stefan Rampp, Hiroshi Otsubo, Yvonne Höller, Joyce Y Wu, Eishi Asano, et al. High-frequency oscillations: the state of clinical research. *Epilepsia*, 58(8):1316–1329, 2017.

[36] Fang Wang, Sheng-hua Zhong, Jianfeng Peng, Jianmin Jiang, and Yan Liu. Data Augmentation for EEG-Based Emotion Recognition with Deep Convolutional Neural Networks. In *MultiMedia Modeling*, volume 10705, pages 82–93. Springer International Publishing, 2018. Series Title: Lecture Notes in Computer Science.

[37] Tarek Lajnef, Sahbi Chaibi, Perrine Ruby, Pierre-Emmanuel Aguera, Jean-Baptiste Eichenlaub, Mounir Samet, Abdennaceur Kachouri, and Karim Jerbi. Learning Machines and Sleeping Brains: Automatic Sleep Stage Classification using Decision-Tree Multi-Class Support Vector Machines. *Journal of Neuroscience Methods*, 250(November):94–105, 2015.

[38] Jobu Watanabe, Motoaki Sugiura, Naoki Miura, Yoshihiko Watanabe, Yasuhiro Maeda, Yoshihiko Matsue, and Ryuta Kawashima. The human parietal cortex is involved in spatial processing of tongue movement-an fMRI study. 21(4):1289–1299, 2004.

[39] Olivier Deiss, Siddharth Biswal, Jing Jin, Haoqi Sun, M. Brandon Westover, and Jimeng Sun. HAMLET: Interpretable Human And Machine co-LEarning Technique. *arXiv:1803.09702*, 2018.

[40] Wilder Penfield and Herbert H Jasper. *Epilepsy and the functional anatomy of the human brain.* 1954.

[41] Hubert Banville, Sean U.N. Wood, Chris Aimone, Denis-Alexander Engemann, and Alexandre Gramfort. Robust learning from corrupted EEG with dynamic spatial filtering. *NeuroImage*, 251:118994, 2022.

[42] François Perrin, Jacques Pernier, Olivier Bertrand, and Jean Francois Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, 72(2):184–187, 1989.