Data Augmentation for Deep-Learning-Based Electroencephalography

Elnaz Lashgari, Dehua Liang, Uri Maoz

PII: S0165-0270(20)30308-3

DOI: https://doi.org/10.1016/j.jneumeth.2020.108885

Reference: NSM 108885

To appear in: *Journal of Neuroscience Methods*

Received Date: 21 February 2020

Revised Date: 10 July 2020

Accepted Date: 24 July 2020

Please cite this article as: Lashgari E, Liang D, Maoz U, Data Augmentation for Deep-Learning-Based Electroencephalography, *Journal of Neuroscience Methods* (2020), doi: https://doi.org/10.1016/j.jneumeth.2020.108885

Data Augmentation for Deep-Learning-Based Electroencephalography

Elnaz Lashgari [1,2]Lashgari@chapman.edu, Dehua Liang [1,2], Uri Maoz [1,2,3,4,5]

[1] Schmid College of Science and Technology, Chapman University

[2] Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University

[3] Crean College of Health and Behavioral Sciences, Chapman University

[4]Anderson School of Management, University of California Los Angeles

[5]Biology and Bioengineering, California Institute of Technology

E-mail:

Highlights

- Data augmentation (DA) is increasingly used with deep learning (DL) on EEG
- It enhances decoding accuracy left unexplained by 29% on average on the datasets we review
- We analyze which specific DA techniques appear to work best for which EEG tasks
- We tested various DA techniques on an open motor-imagery task and compared the accuracy gains to demonstrate the usefulness of DA for DL-based EEG analysis
- We propose guidelines for reporting parameters for different DA techniques

Abstract

-Background

Data augmentation (DA) has recently been demonstrated to achieve considerable performance gains for deep learning (DL)—increased accuracy and stability and reduced overfitting. Some electroencephalography (EEG) tasks suffer from low samples-to-features ratio, severely reducing DL effectiveness. DA with DL thus holds transformative promise for EEG processing, possibly like DL revolutionized computer vision, etc.

-New method

We review trends and approaches to DA for DL in EEG to address: Which DA approaches exist and are common for which EEG tasks? What input features are used? And, what kind of accuracy gain can be expected?

-Results

DA for DL on EEG begun 5 years ago and is steadily used more. We grouped DA techniques (noise addition, generative adversarial networks, sliding windows, sampling, Fourier transform, recombination of segmentation, and others) and EEG tasks (into seizure detection, sleep stages, motor imagery, mental workload, emotion recognition, motor tasks, and visual tasks). DA efficacy across techniques varied considerably. Noise addition and sliding windows provided the highest accuracy boost; mental workload most benefitted from DA. Sliding window, noise

addition, and sampling methods most common for seizure detection, mental workload, and sleep stages, respectively.

-Comparing with existing methods

Percent of decoding accuracy explained by DA beyond unaugmented accuracy varied between 8% for recombination of segmentation and 36% for noise addition and from 14% for motor imagery to 56% for mental workload—29% on average.

-Conclusions

DA increasingly used and considerably improved DL decoding accuracy on EEG. Additional publications—if adhering to our reporting guidelines—will facilitate more detailed analysis.

Keywords: Electroencephalography, Deep learning, Data augmentation, Review

1. Introduction

Electroencephalography (EEG) measures electric fluctuations in the brain. One use of EEG is to measure rhythmic oscillations, which reflect synchronized activity of substantial populations of neurons. Changes in these rhythmic oscillations during cognitive tasks correlate with task conditions, including perceptual, cognitive, motor, emotional, and other functional processes. This renders such task monitoring tractable using EEG [1]. Several reasons make EEG a useful tool for studying neurocognitive processes. First, it captures cognitive dynamics in the time scale at which cognition occurs—tens to hundreds of milliseconds. Second, EEG can directly measure complex patterns of neural activity within small fractions of a second after stimulus onset. Third, the EEG signal is multidimensional, comprising time and frequency, power, and phase, across many electrodes over the scalp. This multidimensionality facilitates specifying and testing hypotheses that are rooted both in neurophysiology and in psychology [2].

Nevertheless, EEG also suffers from several limitations. First, it is an aggregate signal emanating from the aggregated neuronal activity of millions or more cells, which has been transduced through several layers of tissue, fluid, bone, etc. EEG also suffers from low signal-to-noise ratio (SNR) [1-4]. Though various filtering and de-noising techniques strive to decrease the noise in favor of the underlying neural activity. What is more, EEG is a non-stationary signal—its statistics varying over time [1, 5, 6]. This is especially problematic for online, real-time analysis, where it is inherently models that were trained on past neural data that are used to decode present neural activity [7, 8]. Further, for complex machine-learning models, model training time might be lengthy. Hence, not only is it well outside the scope of real-time analysis, necessitating off-line training, but the statistics of the relevant brain activity may change considerably by the time the model is trained. There have been some attempts at adaptive machine-learning techniques to better track the changing statistics of the signal [9-11]. If that is not enough, EEG is generally recorded using tens to hundreds of electrodes recording simultaneously at hundreds or thousands of samples per electrode, whereas a typical dataset, at least in cognitive neuroscience when looking at discrete experimental events, contains only some hundred to a few thousand samples (i.e., experimental trials) at the most. Hence, the initial ratio of samples to features is low in such datasets. Due to the above, classifiers trained on EEG datasets tend to generalize poorly to data recorded at different times, even on the same individual.

This problem is only exacerbated for datasets involving rare events (e.g., sleep transitions, seizures [1]). These result in datasets that are heavily imbalanced between events and non-events. It thus worth noting that methods have been developed to deal with such imbalanced data [12].

Unfortunately, there are additional challenges: inherent variabilities in brain anatomy, head size, cap placement, and dynamics across subjects considerably limit the generalizability of EEG analyses across individuals [1, 13]. In other words, even if a model is well trained on one experimental subject, it would tend to generalize poorly to other subjects. Thus, most EEG classifiers tend to be subject-specific. Yet, even for a single subject, many time-consuming experimental sessions must be gathered to train the machine-learning models well enough to be useful. To overcome some of the above-mentioned limitations, processing pipelines with domain-specific approaches are often used to clean, extract relevant features from, and then classify, EEG data.

Deep Learning (DL) is a subfield of machine learning that focuses on computational models that typically learn hierarchical representations of the input data through successive non-linear transformations—termed neural networks (NN) (because of their superficial resemblance to biological neural networks in the nervous system) [1, 14, 15]. In the past few years, DL has achieved breakthrough accuracies and discovered intricate structures in complex and high-dimensional data such as image classification [16-18], speech recognition [19-21], machine translation, and more [1]. The architecture of the neural networks, their training procedure, regularization, optimization, and hyper-parameter searches are all active research topics in DL, with advances often resulting in dramatic increases in decoding accuracy.

DL typically thrives on problems where (1) there is a lot of data, and (2) The basic unit of information (e.g., a pixel, a letter) has little overall meaning; but potentially complex, hierarchical combinations of such units are useful in understanding the sample. Successful machine-learning classification also at least has the potential to make considerable impact on EEG decoding, remarkably simplifying its processing pipelines for example. It could possibly enable automatic end-to-end learning of preprocessing, feature extraction, and classification modules, while also reaching competitive performance on the target task [22, 23]. However, DL models are typically complex—i.e., have many free parameters (or degrees of freedom) to fit. Thus, if we lack enough data on which to train, training such DL models risks overfitting those models to specific quirks of the training set, limiting the generalizability of the model to an independent test set. One solution to overfitting is therefore to reduce the complexity of the models—e.g., by reducing the number of free parameters or the range in which such parameters can be fit. This is termed "regularization". Another approach is to increase the amount of data on which to train the model—for example using data augmentation, which will be detailed below.

DL has in particular shown some promise for inter-subject generalization [24], which is especially important when only little data is available per subject. Since individual differences in EEG are large (see above), it is not trivial to share classification models across people. We therefore often need to collect new labeled data to train personal models for new users. In some applications, we hope to acquire models for new subjects as fast as possible and reduce the demand for the amount of labeled data. To achieve this goal, transfer learning methods have been shown to be helpful [25]. In transfer learning, existing subjects are sources, and the new

subject is the target. The target data are divided into calibration sessions for training and subsequent sessions for testing. The first stage of the method is source selection, aimed at locating appropriate sources. The second is style transfer mapping, which reduces the EEG differences between the target and each source. We use few labeled data in the calibration sessions to conduct source selection and style transfer. It has been shown that inter-subject transfer learning techniques can help us avoid re-training DL networks, with is often time and energy consuming [26]. However, if the transfer learning ends up with decreased performance or accuracy for the new model, it is called negative transfer. Transfer learning only works if the initial and target problems of both models are similar enough. If the first round of training data required for the new task is too far from the data of the old task, then the trained models might perform worse than expected. And, regardless of how similar developers may think these two sets of training data are, algorithms may not always "agree" with them. No specific standards have thus far been developed for what tasks are related enough to facilitate transfer learning, which makes it challenging to find solutions to negative transfer.

Given the above, a critical question concerning the application of DL to EEG data is therefore "How much EEG data is enough for a desired accuracy level?" Unfortunately, high-grade EEG data collection requires relatively expensive hardware and a lot of participant time. At the same time, access to large, especially clinical, dataset is often limited by privacy and proprietariness concerns. Therefore, large, openly available EEG datasets are uncommon.

Data augmentation (DA) comprises the generation of new samples to augment an existing dataset by transforming existing samples. This holds the promise to increases the accuracy and stability of the classification, at least for EEG data. Exposing the classifiers to varied representations of its training samples makes the model less biased and more invariant and robust to such transformations when attempting to generalize the model to new datasets [27-29]. DA has proven effective in many fields, such as image processing and object recognition. It has even been demonstrated that it can give very deep neural networks a higher accuracy boost than the other standard approaches, which includes various regularization techniques [29] and in particular Dropout [30, 31]. Functional solutions such as dropout regularization, batch normalization, transfer learning, and pretraining have been developed to try to extend DL for application involving smaller datasets [32]. A good survey of regularization methods in DL has been compiled by Kukacka et al. [33]. In contrast to the regularization approach mentioned above, DA approaches overfitting from the root of the problem—the training dataset. This is done under the assumption that more information can be extracted from the original dataset through augmentation. Such augmentation artificially inflate the training dataset size by either data warping or oversampling [32].

New, augmented data is typically generated using two approaches. The first is by applying geometric transformations: translations, rotations, cropping, flipping, scaling, etc. The second is via the addition of noise to the existing training data. Note that increasing the size of the training set also facilitates training more complex models with additional parameters and/or reducing overfitting. However, unlike images, EEG is a collection of very noisy, somewhat correlated (in time and space), non-stationary time-series from different electrodes. And even if feature extraction is performed, geometric transformations are not directly suitable for EEG data because

those may destroy time-domain features [28]. Also, while a human can easily decide whether an augmented dataset (e.g., of cats or other images) still resembles the original class, the same is not true of augmented EEG signals. In other words, correctly labeling augmented datasets can be difficult. Nevertheless, in recent years DA techniques have received widespread attention and achieved appreciable performance boosts when using DL on EEG signals.

We ran a systematic review on DA in EEG and collected all the papers that we were able to find up to and including 2019. The earliest paper we could find that specifically used the name "data augmentation" was in 2015. (Though it should be noted that this technique has been used before 2015 and went by other names, such as oversampling and in particular SMOTE [34-36].(see Methods) And a testament to the growing importance of DA for EEG is that 37 out of 53 papers we found (70%) were from 2018 and 2019 and 21 (40%) were from 2019 alone. This review paper strives to identify trends and highlight available approaches in DA for DL in EEG to address the following critical questions: (1) What DA approaches exist for EEG? (2) Which dataset and EEG classification tasks have been explored with DA? (3) Are there specific DA methods suitable for specific tasks measured by EEG? (4) Which of the input features in EEG are used for training the deep NNs with DA?

2. Methods

2.1. Search method for identification of related studies

The search was conducted on 3rd January 2020 within the Google Scholar, Web of Science, and PubMed databases using the following group of keywords: ('Data Augmentation') AND ('Deep Neural Network' OR 'Deep Learning' OR 'Deep Machine Learning' OR 'Deep Convolutional' OR 'Representation Learning' OR 'Deep Recurrent' OR 'Deep LSTM') AND ('EEG' OR 'Electroencephalography'). Only studies that met the inclusion criteria (Fig. 1) are included below. Further, duplicates among these databases were removed from the search results. The full texts of the remaining studies were then screened.

Inclusion criteria	Exclusion criteria
• EEG classification—This review focused solely on classification based on EEG signals. • Deep learning—In this review, DL is defined as learning using a neural network with at least one hidden layer • EEG augmentation— This review focused on the augmentation of EEG signals. • English Journal and conference papers, as well as electronic preprints, published were chosen as the target of this review. • Studies focusing only on "EEG" AND "DL" and "DA"	• Other studies, such as power analysis and feature selection with no end classification, were excluded. • review papers were excluded

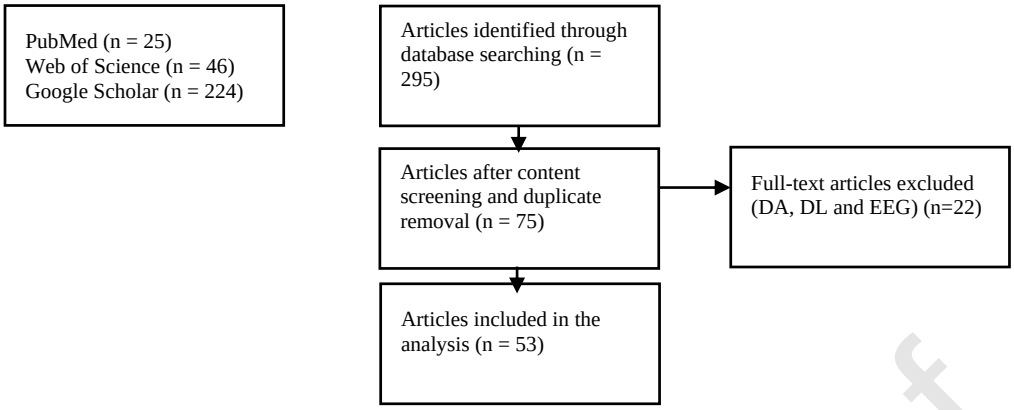

Figure 1 Selection process for the papers

The database queries yielded 295 matching results. Of those, 32 were duplicated. Manually screening the remaining 263 papers suggested that 188 of them were not relevant for this review (e.g., the keywords were included in the references rather than in the paper itself). We thus ended up with 75 papers, which we read carefully to make sure they meet all our inclusion criteria. We found 22 that did not meet the inclusion criteria following closer inspection (e.g., they did not focus on human subject, they did not include classification, and so on.) Hence, based on our inclusion and exclusion criteria, 53 papers were selected for inclusion in this analysis (Figure 1**)**. The earliest one was from 2015.

Regarding our inclusion criteria, we should also mention more specifically that the DA and NN as search terms were found before 2015. For instance Image augmentation in the form of data warping can be found in LeNet-5 (1998) [37]. This was one of the first applications of CNNs on handwritten digit classification. Data augmentation has also been investigated in oversampling applications. Oversampling is a technique used to re-sample imbalanced class distributions such that the model is not overly biased towards labeling instances as the majority class type. Oversampling was applied on EEG signal since 1970s [38] [39]. However, none of these papers meet our inclusion criteria.

2.2. Data Extraction and presentation

For each selected paper, around 40 features were extracted covering 7 categories: Origin of the article, DA types, Dataset, Task information, Preprocessing, DL strategy, Results (Table 1).

Table 1. Data items extracted for each article selected

Category	Data item
Article origin	Type of publication (Journal article, conference article, or in an electronic preprint repository)
Data augmentation (DA)	DA technique used to generate new samples Parameters for DA Magnification factor (m)
Dataset	Quantity of data, subjects, classes, channels

Task information	Task type
Preprocessing	Frequency range used for analysis EEG signal features
Deep-learning strategy	Main characteristics of NN, such as number of convolutional layers, hidden layers, activation function of hidden layers and output.
Results	Decoding accuracy

3. Results

3.1. Origin of the selected studies

Our research methodology returned 26 journal papers,16 conference and workshop papers, and 11 preprints (arXiv or bioRxiv) that met our inclusion criteria. There were 4 papers in IEEE Transactions on Neural Systems and Rehabilitation Engineering, 3 papers in Biomedical Signal Processing and Control and the rest of the papers were each in a different journal (see Table 1 for details). Interestingly, we found no papers that fulfilled our search criteria before 2015. Further, testament to the growing importance of DA for EEG is the clear year-by-year rise in the number of papers answering our search criteria from 2015 to 2019 (Figure 2A).

3.2. EEG classification task

The EEG tasks in these papers fell into 7 groups: seizure-detection (24%), motor imagery (21%), sleep stages (15%), emotion recognition (15%), mental workload (9%), motor task (8%), and visual task (8%) (Figure 2B). The following describes these EEG tasks (see Table 6 for more details):

3.2.1. Seizure-detection studies. A seizure is a sudden, uncontrolled disturbance in the electrical activity of the brain. For seizure detection in epilepsy, EEG signals are recorded during seizure and non-seizure periods. The goal of these studies is to detect upcoming seizure and preemptive notification to the patients [40, 41]. Seizure manifestations on EEG are extremely variable both inter- and intra-patient. Naturally, non-seizure events are easy enough to record. But seizures tend to be rare. DA has been successful at increasing the number of rare events (seizures) in the dataset and thus at increasing the accuracy of seizure-detection algorithms.

3.2.2. Motor imagery tasks. These studies instruct subjects to imagine moving their limbs, tongue, or other body parts. Motor imagery EEG decoding is an important method in brain-computer interfaces (BCI) that has the potential to help highly disabled people communicate with the outside world without relying on muscle activity (e.g. [42]).

3.2.3. Sleep stages scoring tasks. Studies on sleep-stage classification record the EEG signal of subjects overnight. These signals are then scored and classified to wakefulness (W) and then 4 stages of sleep based on the American Academy of Sleep Medicine(AASM) scoring manual: Rapid eye movements, or REM (R) and 3 non-REM stages (N1, N2, and N3) [43, 44]. The eventual application of this research focuses on sleep related disorders, such as sleep apnea, insomnia, and narcolepsy (e.g. [45, 46]).

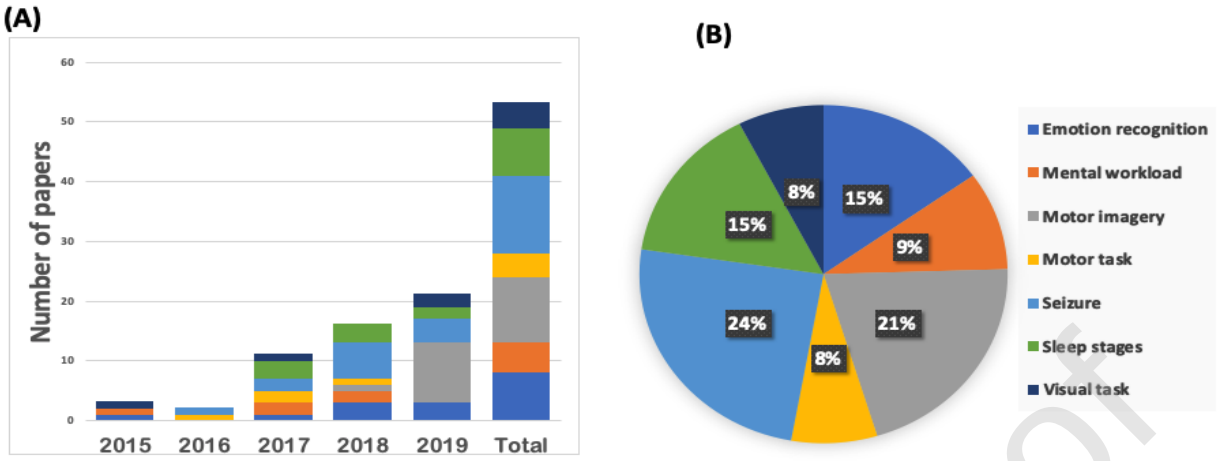

Figure 2. EEG classification task. (A) Number of publications per domain of EEG task per year. (B)The percentage of different EEG classification task across all studies.

3.2.4. Emotion recognition tasks. Here subjects watch video clips, which have been categorized by experts as eliciting various emotions. Facial expressions and EEG signals are then recorded from the subjects. However, some subjects may hide their real emotions using misleading facial expressions. Therefore, EEG signals and emotion self-assessment typically follows. The result can be parsed into valence and arousal scales. Emotion recognition is a crucial problem in human-computer interaction (HCI) for example: virtual reality, video games, and educational systems (e.g. [28]).

3.2.5. Mental workload tasks. Subjects are here instructed to carry out different mental tasks of varying complexity. The results of these studies reflect the interaction between the human inner cognitive capacity and the level of task complexity. Research into mental workload has applications in BCI performance monitoring and in cognitive stress monitoring (e.g. [47]).

3.2.6. Motor tasks. Here, subjects are instructed to either rest or move some parts of their bodies. Researchers use such tasks to design, modify or improve classification methods for different applications (e.g. [48]).**3.2.7. Visual tasks.** These studies focus on the detection and classification of the intentions and decisions of subjects while they watch rapidly changing sequences of pictures or letters. This helps to improved non-verbal communication systems and BCI (e.g. [49]).

3.3. Data and Reproducibility

We collected dataset information for 53 papers. This information included:

- Data quantity: Amount of data in the study (total hours of recording or number of samples)
- Number of Channels: Number of channels recorded and which of them were used for analysis
- Subjects: Number of recorded participants and which of them were analyzed
- Dataset: Publicly available, proprietary, etc.

See Table 6 for more details.

3.4. Pre-processing and Feature extraction

The analysis of EEG signals is typically carried out by one of two methods. The first is event-related potentials, which are fluctuations of the potentials over time that are locked to an event (e.g., to 'stimulus onset' or 'button press'). The second is spectral analysis of rhythmic oscillations, which reflect the synchronized activity of very large populations of neurons. Regardless of the analysis method, it is an aggregate signal emanating from the neuronal activity of millions or more brain cells, which has been transduced through several layers of tissue, fluid, bone, etc. It also potentially includes undesired electrophysiological signals, such as electromyograms (EMG) of muscle contractions specifically eye blinks, heart beats, and others. Therefore, the EEG signal is inherently noisy. Though various filtering and de-noising techniques strive to decrease the noise in favor of the underlying neural activity. In the 53 studies we found, 85% (45 studies) removed the artifacts manually—mainly using high, low, and band pass filtering. Importantly, this means that 15% of the studies (8 studies) did not remove artifacts manually. Of those 8 studies, 7 did not take any action to remove artifacts, and the remaining study did not address artifact removal.

Most studies used frequency domain filters to limit the bandwidth of the EEG signals. This enabled them to focus on a certain frequency range that was of interest. Roughly, half of the reviewed papers low pass filtered the signal below low gamma band or 40 Hz. The filtered frequency ranges organized by task type (Figure 3). We found that there were no studies that specifically check the role of this filtering for NN [23].

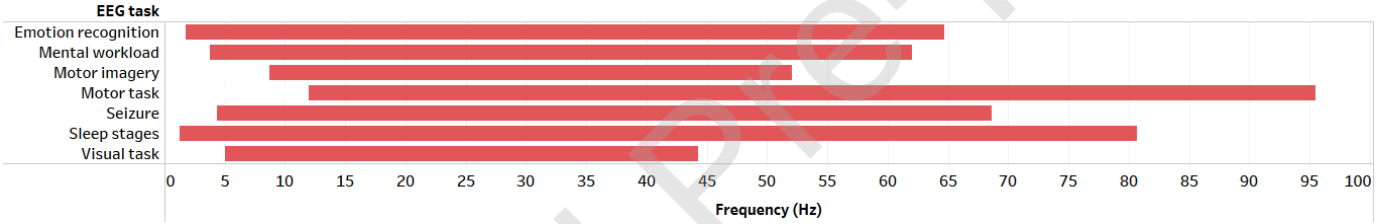

Figure 3. Frequency range used in EEG analysis for each identified study, organized by EEG task type.

3.4.1. Input Formulation

The inputs to the NNs in the studies that fulfilled our inclusion criteria, fell into three categories. The first included raw EEG signals (in the time domain) (36%). The second calculated features from the raw signals and used those as inputs (49%). And the third used spectrograms, processed as images (15%). The selection of input formulation heavily depended on the task type and deep-learning architecture. Thus, we can see that most of the studies used calculated features to train their proposed NNs. When attempting to find behavioral patterns, it is common to analyze specific frequency ranges of EEG signals. Wavelet, entropy, spatial filter, short-time Fourier transform (STFT), spatio-temporal features, and power spectral density were used in the reviewed papers to calculate the features of EEG the signals. Raw EEG values was another popular feature for training NN. It's interesting that NNs can learn complicated features from large amount of raw data. Many NNs, especially RNN, used spectrogram and fast Fourier transform (FFT) to convert EEG signals to images (Figure 4). Hence, in general, studies run the gamut from using raw EEG to heavily engineered features. When we analyzed the studies that fulfilled our inclusion criteria based on the input formulation and on the EEG task, we found that (Emotion recognition, mental workload, motor imagery, and seizure) mostly used calculated

features. Motor task, sleep stages, and visual task chose signal values as their input primarily (Figure 4).

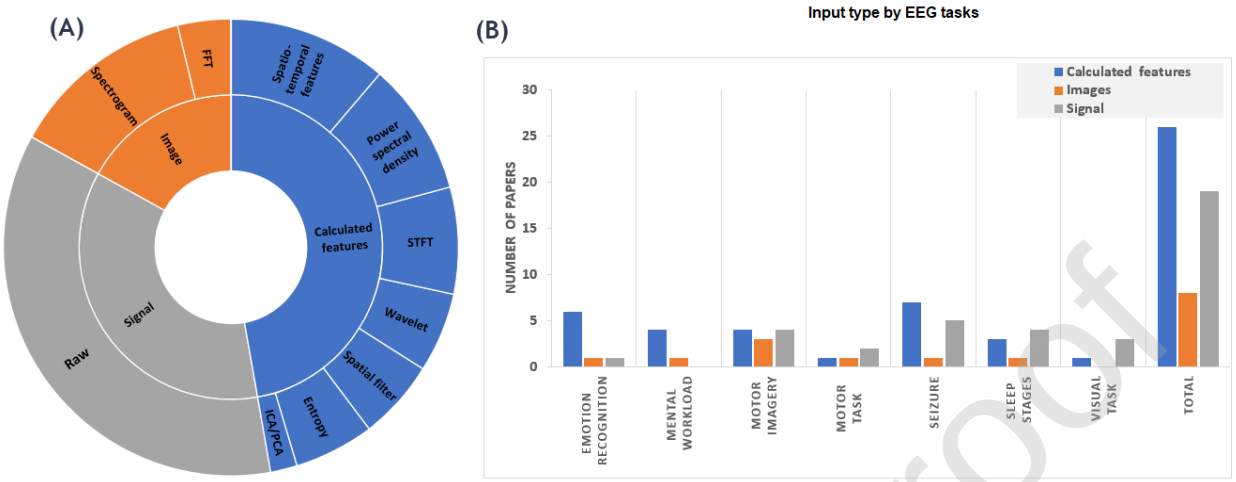

Figure 4. Input formulation across all reviewed papers. (A) The inner circle shows the general input formulation, while the outer circle shows more specific details. (B) Number of papers for general input formulation compared across different tasks.

3.5. Deep learning architectures

Deep learning is a subfield of machine learning based on artificial neural networks, which can be thought of as learn hierarchical representations of the input data through non-linear transformations. While beginning to rise to prominence in the late 2000's, in the few years since, it has arguably revolutionized the field, achieving remarkable accuracy on and discovering intricate structures in complex and high-dimensional data, such as image classification, speech recognition, and automated translation. Various deep learning architectures have been developed since, with this fast-moving research field routinely producing new architectures. We discerned 6 different categories in deep learning: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Multi-layer perceptrons (MLP), Stacked Auto Encoders (SAE), Long Short-Term Memory (LSTM), and hybrid combinations of the above. By order of prevalence these were: CNN (62%), Hybrid (16%), MLP (8%), SAE (6%), LSTM (6%), and RNN (2%) (Fig. 5).

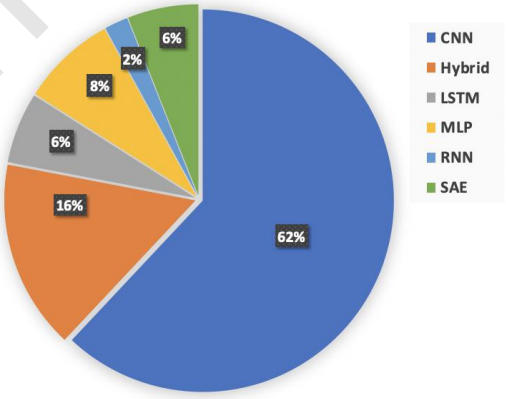

Figure 5. Deep learning architecture across all studies

Figure 6 visualizes the aggregated information about DL architecture of reviewed studies. This figure helps to understanding the trends in the formation of specific deep-learning architectures. For more details see Table 6.

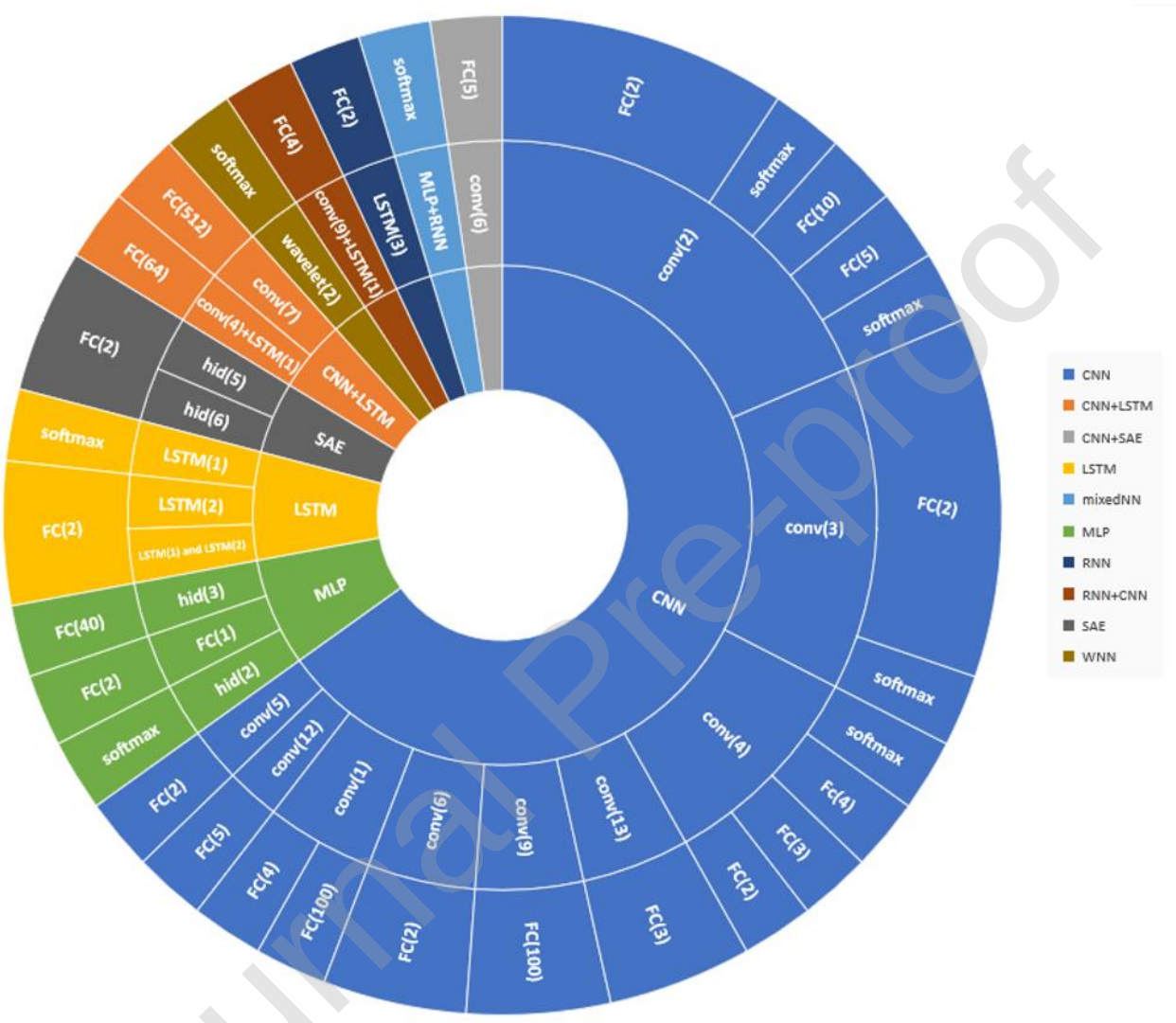

Figure 6. Aggregated information of deep learning architectures. The inner circle shows the general DL architecture, the middle circle, shows the primary design features, such as the hidden layers or convolutional layers, and the outer circle shows the last layer of DL architecture. FC: Fully connected, hid: Hidden layers, softmax: Softmax fuction.

Figure 7, visualizes the proportion of input formulation by DL architecture. As is apparent, the specific input formulation strategies varied significantly as a function of the type of the deep learning architecture. While there was not a clear consensus for all studies together, RNN and SAE architectures used only images and calculated features as inputs, respectively. Hybrid, CNN

and MLP studies included instances of all 3 information types. Interestingly MLP and CNN used directly signal values as inputs.

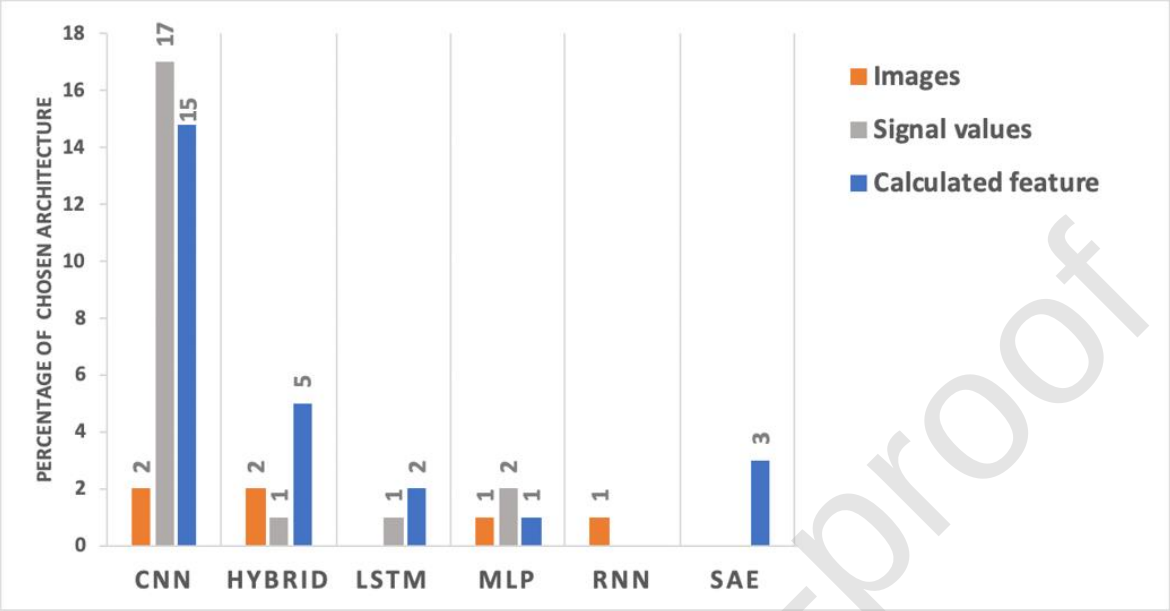

Figure 7. The percentage of input formulation by chosen DL architecture

3.6. Data Augmentation methods

This section details the methods found for methods that have so far been used to augment EEG signal for machine learning. Data augmentation (DA) comprises the generation of new samples to augment an existing dataset by transforming the existing samples in a manner that increases the accuracy and stability of the classification or regression. Exposing the classifier to more variable representations of its training samples makes the model more invariant and robust to transformations of the type that it is likely to encounter when attempting to generalize to unseen samples. Further, increasing the size of the training set facilitates training more complex models with additional parameters and/or reducing overfitting. In recent years, DA techniques have received widespread attention and achieved appreciable performance boosts for DL on EEG signals. Here we cover all the papers that we were able to find up to and including 2019. The first paper was found in 2015. The testament to the growing importance of DA for EEG is that 37 out of 53 papers (72%) we found are from 2018 and 2019 Figure 8.

The DA for DL-based EEG in 53 papers fell into 7 categories in our analysis: noise addition (17%), GAN (21%), sliding window (24%), sampling (17%), Fourier transform (4%), recombination of segmentation (6%) and other (11%) Figure 8. Below we discuss each DA method in much more detail.

3.6.1. Noise addition

In our research, we found two main categories for adding noise to the EEG signals in purpose of DA: (1) Add various types of noise such as Gaussian, Poisson, Salt and pepper noise, etc. with different parameters (for instance: mean ($\boldsymbol{\mu}$) and standard deviation ($\boldsymbol{\sigma}$)) to the raw signal (2) Convert EEG signals to sequences of images and add noise to the images. Nine papers used noise addition method to increase training dataset.

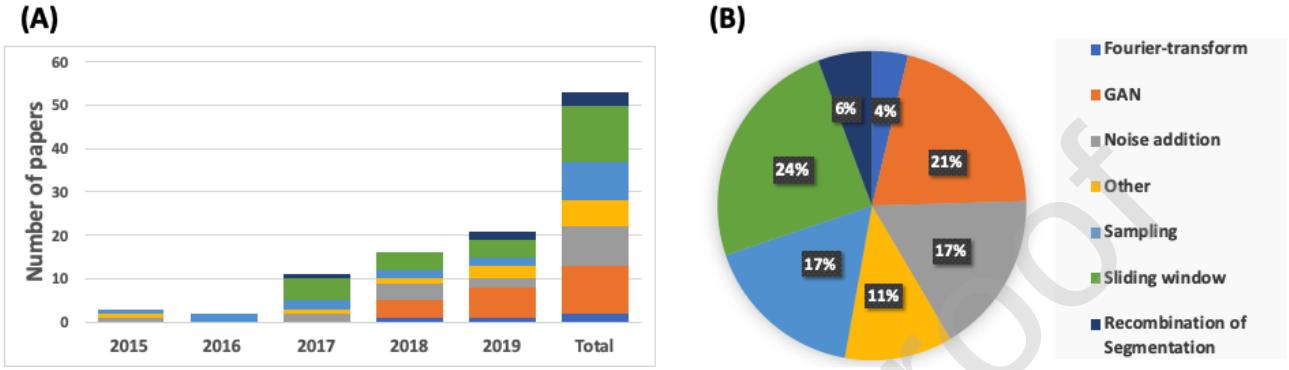

Figure 8. DA across all studies. (A) Number of publications per domain of DA per year. (B) The percentage of different DA methods across all studies. Note that we only collected data until January 2020.

In 2015, Bashivan et al. transformed EEG signals into a sequence of topology-preserving multi-spectral (2D feature images) in a specific time interval **[50]**. Fast Fourier Transform (FFT) was performed on the mental load EEG signals to estimate the power spectrum of the signal in three frequency bands of theta (4-7Hz), alpha (8,13Hz), and beta (13-30Hz). A single image was constructed from spectral power within three prominent frequency band which is extracted from each electrode location. The sequences of image representations fed into the LSTM and CNN for the EEG classification. For addressing the unbalanced ratio between number of samples and number of model parameters, they randomly added various noise level to the images. However, augmenting the dataset did not improve the classification performance and even for higher value of noise, the error rate increased.

Z. Yin et al. (2017) proposed an adaptive DL model based on Stacked Denoising AutoEncoders (SDAE), which was designed for cross-session Mental Workload (MW) classification using EEG [51, 52]. They could increase the accuracy of their model by adding Gaussian white noise to the EEG feature vector ($\mu = 0.01, m = 2,3,4,5,6$). This vector contains centroid frequency, log-energy entropy, mean, five power components, Shannon entropy, sum of energy, variance, zero-crossing rate of each channel and power differences between four selected channel pairs. Their classification accuracy on an independent dataset improved from 76.5% (without DA) to 85.5% (with DA). The highest classification accuracy was achieved with m=6 and the lowest with m=0 (without DA). They concluded that the number of samples (trials) in the original dataset was insufficient for training the NN.

Wang et al. (2018) added Gaussian white noise to their training data (in the time domain) to obtain new samples for an emotion-recognition task [28]. In their experiments, EEG signals were recorded while subjects were watching emotionally loaded videos. They used differential entropy (DE) features to train their proposed classifiers. For EEG signals, the DE feature is equivalent to the logarithm of the energy spectrum in the delta (1–3 Hz), theta (4–7 Hz), alpha (8–13 Hz), beta

(14–30 Hz), and gamma (31–50 Hz) frequency bands. The authors opted for Gaussian noise due to concerns that adding some local noise (i.e., noise that affects EEG data locally) such as Poisson or salt-and-pepper may change the intrinsic features of EEG signals. The experimental results on SEED dataset showed that by augmenting training dataset 30 times, the accuracy of ResNet improved from 34.2% to 75%, better than LeNet (from 49.6% to 74.3%).

R. Hussein et al. (2018) used another DL technique, using a recurrent neural network (RNN) and Long Short-Term Memory (LSTM) network. Their goal was automatic detection of epileptic seizures using EEG signals [53]. And they reported that they improved the robustness of their model by adding Gaussian white noise, muscle artifacts and eye-blinking. Though they did not give any specific details about white noise they used as DA methods and the model performance wasn't compared for DA vs. non-DA.

S. Kuanar et al. (2018) also used an LSTM with a convolutional neural network (CNN) to learn robust features and predict the levels of cognitive load from EEG recordings [47]. They transformed the EEG time-series into a sequence of multispectral images that carried spatial information—theta (4-7Hz), alpha (8-13Hz), and beta (13-30 Hz). The data was once again augmented by adding various Gaussian noise level to the images. Though they did not give any specific details about white noise they used as DA methods and the model performance wasn't compared for DA vs. non-DA.

E. Salama et al. (2018) generated the noisy EEG signals, by adding Gaussian noise with zero mean and unit variance to the original input EEG training dataset [54]. They set the signal-to-noise ratio (SNR) between original EEG signal and the noisy to 5. The DA phase enhanced the performance of the proposed 3D-CNN on emotion recognition dataset. For valence and arousal classification, they achieved 79.11%(without DA) and 88.49%(with DA). For 4 combinations of valence and arousal —(low valence-low arousal), (low valence-high arousal),(high valence-low arousal) and (high valence-high arousal) they obtained 79.11%(without DA) and 87.44%(with DA).

Parvan et al. (2019) doubled the number of trials of BCI competition IV dataset 2b by adding gaussian noise with zero mean and a standard deviation of 0.15 to ovoid overfitting [55]. Their proposed CNN had 4 convolution layers as well as data augmentation and resulted in a 0.07 improvemnet in the kappa coefficient [55].

Y. Li et al. (2019) emphasized the fact that increasing depth of CNN causes a higher classification accuracy. However, doing so may aggravate the vanishing-gradient problem and substantially increase the number of trainable parameters to be tuned, and these models may tend to be overfitting easily [56]. For four-class motor imagery task, they exploit the standard deviation of Gaussian noise in the DA affects the classification result. The optimal standard deviation is 0.001with zero mean on 2 imagery task datasets (table). It is noticeable that for almost all subjects, the performance has been significantly improved after DA. Furthermore, by comparing confusion matrix before and after DA, they showed that for a specific imagery task, DA worked well except for one task(feet). Table 2 shows all the papers used noise addition as their DA technique. From this table, we can see that there is lack of information about noise addition parameters ($\boldsymbol{\mu}$: mean, $\boldsymbol{\sigma}$: standard deviation), magnification factor (m) and reported

accuracy before and after DA. Maybe this is because that their problem wasn't DA topic and they wanted to increase just performance accuracy.

Table 2. All reviewed papers that used noise addition as their DA technique

Study	Dataset	Task information	Input formulation	Deep learning strategy	Noise Addition parameters	Accuracy (without DA)	Accuracy (with DA)
[50]	University of Memphis Institutional review board 13 subjects 2670 trials 64 channels	Mental workload 4-13 Hz	Images, FFT	CNN+LSTM Conv(7) + FC(512) Relu, softmax	NA	NA	Did not improved
[51]	SDAE 8 subjects 180 min/subject 1 channel 2 class	Mental workload 1.5-40Hz	Calculated features, Power spectral density	SAE Hid(5) + FC(2) Sigmoid, NA	$\sigma = 0.01$, $\mu = 0, m = 6$	NA	93%
[52]	AutoCAM 7 subjects 1h 11 channel	Mental workload 1-40 Hz	Calculated features, FFT and power spectral	SAE Hid(6) + FC(2) Sigmoid, sigmoid	$\sigma = [0.1,0.2, \dots 1.5]$, $\mu = 0$, $m = 6$	76.5%	85.5%
[28]	SEED 14 subjects 1890 trials 62 channels 3 class	Emotional recognition 1-50 Hz	Calculated features, Entropy	CNN Conv(4) + FC(3) sigmoid	$\sigma = 0.2$, $\mu = 0, m = 30$	49.6%	74.3%
[28]	SEED 14 subjects 1890 trials 62 channels 3 class	Emotional recognition 1-50 Hz	Calculated features, Entropy	CNN Conv(13)+FC(3) sigmoid	$\sigma = 0.2$, $\mu = 0, m = 30$	34.2%	75%
[28]	MAHNOB-HCI 30 subjects 527 trials	Emotional recognition 1-50Hz	Calculated features, Entropy	CNN Conv(13)+FC(3)	$\sigma = 0.2$, $\mu = 0, m = 30$	40.8%	45.4%

	32 channel 3 class						
[53]	Bonn University 5 subject [2,3,5] class	Seizure [0.53, 40] Hz	Raw signal	LSTM softmax	Gaussian white noise+(muscle and eye blink)	NA	2 class: 99%
[47]	NIMHANS 22 subject, 6490 trials(8 hours) 64 channels 4 class	Mental workload 4-30 Hz	Calculated features, power spectral density	RNN+CNN Conv(9)+LSTM(1) Relu, softmax	NA Add noise to image M: NA	NA	93%
[54]	DEAP 32 subject 40min/subject 32channels 2 and 4 class	Emotion recognition [1,50]U[60, end) Hz	Calculated features Spatio-temporal	CNN Conv(2) Relu, softmax	$\sigma = 1, \quad \mu = 0,$ m=[10,30,50]	79.11 % 79.12 %	88.49 % 87.44 %
[55]	BCI competition IV 2b 9 subjects 5 sessions 3 channels	Motor Imagery [0.5,100] Hz	Raw signal	CNN Conv(4)+FC(2) Elu, softmax	$\sigma = 0.15$, $\mu = 0$, $m = 2$	NA	NA
[56]	BCI competition IV 2a 9 subject 72trials/subject 22 channel 4 class	Motor imagery 7-125 Hz	Calculated features Spatio-temporal	CNN Conv(1) FC(4)+softmax Relu, sigmoid	$\sigma = 0.001,$ $\mu = 0,$	Reported subject by subject e.g. 70%	Increased 77.9%
[56]	High Gamma dataset(HGD) 30 subjects 7000trials/subject 1channel 2 class	Motor imagery 7-125 Hz	Calculated features Spatio-temporal	CNN Conv(1) FC(4)+softmax Relu, sigmoid	$\sigma = 0.001,$ $\mu = 0,$	NA	NA

3.6.2. Generative adversarial network

The term Generative Adversarial Network (GAN) was first demonstrated by Goodfellow, et al. as a new framework to learn the underlying distribution of data from two competing networks: the generator (G) and the discriminator (D). While the generator makes "fake data", the discriminator classifies the "fake data" as real or fake using the given label as if they were playing a minimax game [57].

During the process, the generator gets better at generating data that are similar to the real data, until the discriminator fails to distinguish real from fake data Figure 9. The minimax game of a GAN is given by:

$$\min_D \max_G V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[1 - \log D(G(z))],$$

where p_{data} is the distribution of the real data and p_z is gaussian noise. $D(x)$ gives a probability of an input x belonging to the real data, while $G(z)$ produces fake samples that strive to trick D by learning how to produce data that appears to come from the distribution of the real samples, p_{data}. The optimization process utilized the Jensen-Shannon (JS) Divergence to find the minimum of the function [57].

GANs have been widely applied for generating data in many disciplines outside neuroscience and EEG. For example, In the method that Zhang et al. (2017) proposed a GAN was used to generate images from text [58]. Bousmalis, K., et al. (2017) strives to generate rendered images that are similar to images in a dataset [59]. Antoniou et al. (2017)used a GAN to create new data from three different popular image datasets: Omniglot, EMNIST, and VGG-face [60].

Specifically for augmenting EEG signals, Zhang et al. (2018) proposed a conditional deep convolutional generative adversarial network (cDCGAN) [61]. The cDCGAN is an improved version of the GAN that uses information from the labels and adds them to the model as conditional properties:

$$\min_D \max_G V(D,G) = E_{x\sim p_{data}(x)}[\log D(x|y_z)] + E_{z\sim p_z(z)}\big[1 - \log D\big(G(z|y_d)\big)\big],$$

where y_z and y_d is the information from the corresponding labels. The dataset contained EEG signals recorded over 3 electrodes, and composed of 7 sessions, with 40 trials per second, each lasting 9 seconds. It was collected while subjects were asked to imagine moving either left or right. A CNN was trained to classify each EEG signal as Left or Right.

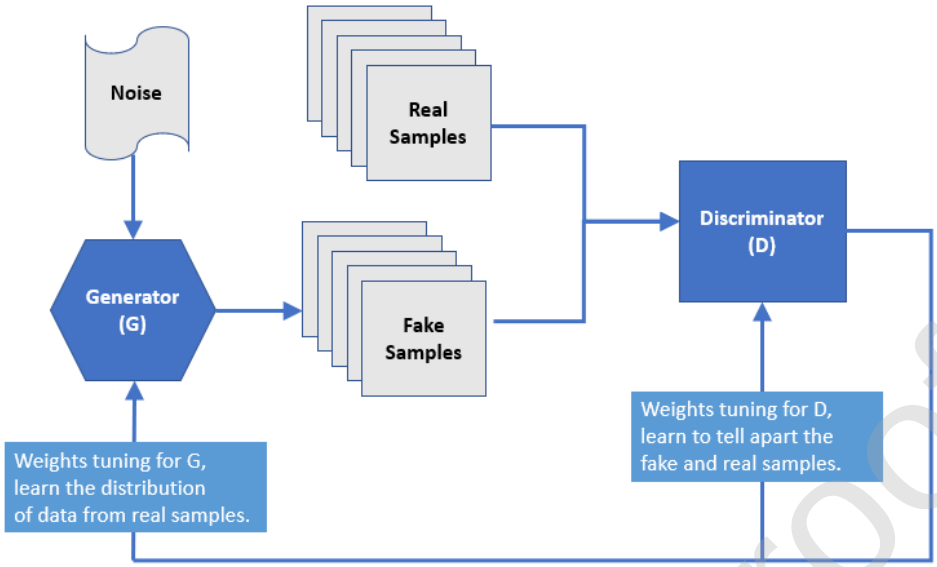

Figure 9 Diagram of Generative Adversarial Network

The EEG signals were preprocessed before feeding them into the CNN. Only 5 out of the 9 seconds of EEG in each trial were selected for processing and only alpha (7-15 Hz) frequency components were extracted as time-frequency features. Using data generated from the cDCGAN, classification accuracy increased from 83% to 86%. The authors compare the accuracies for models trained using different proportions of artificial data. However, the largest dataset only doubles the original dataset in the experiment (i.e. m=2), while others have used larger augmentation.

Piplani et al. (2018) used a GAN to generate more EEG data to increase the robustness of a "passthought authentication system" that uses the user's EEG signals to securely log into devices [62]. The EEG signals were collected using a device with only a single channel at a sampling rate of 500Hz. The 'negative' samples were collected from 30 subjects who were asked to perform a series of mental task for 5 minutes while EEG was recorded. The 'positive' samples were collected from one subject while the subject was doing the same mental tasks for 5 mins and free to do any tasks for another 5 minutes. The dataset that trains the selected model, XGBoost, consists of 30,000 negative samples and 40,000 positive samples. Each sample is a segment of the EEG signal. These data were augmented with 10,000 artificial EEG signals that were generated from a GAN. This increased the accuracy of the model from 90.8% to 95.0%, which is noteworthy for such high accuracies.

Zhang et al. (2018) proposed a framework called Deep Adversarial Data Augmentation (DADA) for generating new data, allowing deep network classifiers to be trained on small datasets [63]. They further investigated and compared different traditional approaches for dealing with small datasets in DL applications—such as dimensionality reduction, semi-supervised learning, transfer learning, and data augmentation. DA was widely used for image data because images can be altered easily—maintaining their content on the one hand while increasing the variance of the representation of that content by rotating, cropping, scaling or just adding noise to the

original dataset. However, these techniques are usually not suitable for non-image data such as EEG signals. One of the examples in this study focuses on increasing the size of an EEG dataset from a BCI competition [64]. This dataset contained 3 channels (C3, Cz, and C4) of EEG collected from 400 trials of motor imaginary tasks. Time-frequency features were extracted from these EEG signals, which formed a 32 x 32 x 3 image for each EEG signal, which was in turn used for training a CNN classifier. Compared to the traditional GAN, DADA was able to generate more diverse artificial data because of its redesigned loss function. A traditional GAN trains the discriminator on only 2 classes. In contrast, DADA uses 2K classes for the discriminator, K for each of the real and artificial datasets. They found that accuracy increased from 74.8%, using a traditional CNN as a benchmark, to 79.3%, using the DADA model.

Hartmann et al. (2018) used a slightly modified version of the Wasserstein Generative Adversarial Network (WGAN) to generate new EEG signals [65]. Training an original GAN suffered from vanishing gradients while optimizing the JS Divergence [57]. A WGAN solved this problem by minimizing the Wasserstein distance:

$$W(p_{data}, p_{fake}) = E_{x\sim p_{data}}[D(x)] - E_{x\sim p_{fake}}[D(x)],$$

where p_{fake} is the distribution of the generator that generates fake (or artificial) samples. In addition, a gradient penalty term $P(p_{\hat{x}}) = \lambda \cdot E_{\hat{x}\sim p_x}[max(0, \|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]$ was also added to produce a useful gradient, where $p_{\hat{x}}$ is the distribution of $\hat{x}$ that are points on a line connecting the real and fake data. Hartmann et al. improved the model by scaling λ, allowing the parameter to adjust its impact based on different Wasserstein distances [65]:

$$L = -W(p_{data}, p_{fake}) + \max\left(W(p_{data}, p_{fake})\right) \cdot P(p_{\hat{x}})$$

The EEG signals were collected from a simple motor task experiment, in which subjects were asked to raise their left hand or to rest. There were 438 trials in total—286 were used for training, 72 for validation, and 80 for testing. Only one channel, FCC4h, was included in this experiment. All total 438 signals were used to train the WGAN model. Unlike other studies that only used classification accuracies to compare the quality of new generated data and the original data, four other evaluation metrics were used in this study: the inception score (IS) [66], Frechest inception distance (FID) [67], Euclidean distance (ED), and sliced Wasserstein distance (SWD) [68]. After comparing the different metrics, optimizing the GAN for good IS and FID produced the best EEG data approximations [65]. This method did not use any classification algorithms to validate the accuracy, therefore it is not included in Table 3 for accuracy comparison.

A conditional version of the WGAN was used by Luo and Lu (2018) to augment EEG data. Similar to cDCGAN, WGAN also utilized the label information to infer the distribution of the real data [69]. The datasets used to test the WGAN model were SEED [70] and DEAP [71]; two popular public EEG datasets for emotion recognition. The EEG signals from the SEED dataset had 62 channels. They were collected from 15 subjects while they were watching film clips selected to induce positive, negative, or neutral emotions. For each subject, 3394 epochs were recorded. The DEAP dataset had 32 channels of EEG signals recorded from 32 subjects, with 2400 epochs each while they were watching music videos. There were 2 classification tasks for the DEAP dataset: high vs low arousal and high vs low valence. Luo and Lu tried different sizes

for the augmented data and found that doubling the data (m=2) provided the highest accuracy comparing to other attempts (m=0.5, 1.0, 1.5). An SVM classifier trained on the augmented dataset improved 2.97% for the SEED dataset from 83.99% to 86.96%. DA seemed to have a larger effect on the DEAP dataset. While classifying arousal, there was a 9.15% improvement in classification accuracy from 69.02% to 78.71%. For valence classification, the improvement was even larger with a 20.13% increase from 53.76% to 73.89%. The method did not specifically mention the chance level accuracy for both datasets. For the SEED dataset, since there are three classes, we are assuming that the chance level accuracy is 33.33%. For DEAP dataset, the chance level accuracy is 50% for binary classification.

In 2019, Luo et al. adopted a conditional Boundary Equilibrium GAN (cBEGAN) to generate artificial differential entropy features of EEG signals on 2 popular emotion recognition dataset (SEED, SEED V) [72]. cBEGAN used the Wasserstein distance to measure the difference between two reconstruction loss distributions. The main advantage of cBEGAN is that it can overcome the instability of conventional GAN and has very quick convergence speed. They generated 50 to 2000 artifacts samples and added them to the original training dataset. With 2000 added samples, the accuracy increases from 81.9% to 87.56% for SEED; and with 1000 samples, the accuracy increases from 54.3% to 62.8% for SEED V, respectively.

Wei et al. (2019) used WGAN with gradient penalty to increase the sample diversity in seizure detection in the CHB-MIT Scalp EEG database (with 23 subjects) [73]. Testing the performance on one patient, they used generated data from the other 22 patients involved in the training. They employed a 12-layers CNN and achieved 81% accuracy (without DA) and 84% (with DA).

Chang et al. (2019) used GAN to increase the size of dataset for a 2-class emotion recognition task [74]. The generator and discriminator of the GAN consists of three hidden layers, which consists of 50, 100, and 50 nodes, respectively. The number of nodes in each layer was determined after evaluations with multiple combinations of hyper parameters that showed the highest training speeds. The generator received random values between 0 and 1 and generated virtual EEG data. The discriminator received EEG collected through experiments and virtual data and distinguished the original data from the virtual data. Once the training was complete, the EEG data generated by the generator were saved. The authors increased the number of trials from 32,000 to 92,000 and by that raised the final accuracy from 97.9% to 98.4%.

Yang et at. (2019) augmentated dataset 2b competition IV BCI using a GAN network [75]. They used CNN-LSTM to classify left and right hand motor imagery task. The average accuracy for 9 subjects was 76.4%. Unfortunately, they did not report the results without DA.

Panwar et al. (2019) proposed using a class conditioned Wasserstein Generative Adversarial Network with gradient penalty (cWGAN-GP) to generate synthetic EEG data of a single channel [76]. The study claims that the cWGAN-GP method is able to counter instability and frequency artifacts problems while training an ordinary GAN [65]. The Wasserstein distance and the gradient penalty stabilized the training process [77]. The class conditioned implementation allowed the generator and discriminator to avoid mode collapsing, which is responsible for trapping the data generated from the GAN in some specific modes [65]. The proposed

architecture had two fully connected layers and two convolutional layers for the generatoras well as three convolutional layers and two fully connected layers for the discriminator. The dataset that the paper used to train the cWGAN-GP was collected during the BCIT X2 Rapid Series Visual Presentation (RSVP) experiment, where subjects were asked to identify target images in an image stream presented at 5Hz [78]. The dataset contained EEG signals from 10 subjects, with 5 sessions and 1 hour of recording per session using a 256-channel BioSemi system. It had two classes, target and non-target, 967 samples each, which were pre-processed using the PREP pipeline [3]. The pipeline performed band-pass filtering from 0.1 to 55Hz, referencing, bad channel interpolation and baselining. One second of signal from each trial after image onset was extracted, down sampled to 64Hz and normalized using the mean and standard deviation from each epoch. The paper used three different methods to evaluate the performance of the data generated from the cWGAN-GP: visual inspection, log-likelihood distance from Gaussian mixture models (GMMs), and classifier performance. The visual inspection and GMM results both showed that the generated data was of high quality. In classifier performance evaluation, the synthetic data size was 3828, and it was added to the training dataset during training. The classifier trained with the synthetic data shows an improvement of 5.18% (from 50.02% to 55.2%) on cross subject evaluation and 3.12% (from 60.8% to 64.08%) on same subject evaluation using a CNN with 3 convolutional layers.

Aznan et al. (2019) had subjects look at one of three different objects, each flickering at 10, 12, or 15 Hz (each at a different frequency). Their goal was to detect which object the subject was looking at using BCI technolgy and then direct a humanoid robot toward that object. They compared three different methods: Deep Convolutional Generative Adversarial Network (DCGAN) [79], gradient panelized Wasserstein Generative Adversarial Network (WGAN-GP) [77], and Variational Auto-encoder (VAE) [80]. They then used those methods to generate synthetic EEG data to improve the classification accuracy on their Steady State Visual Evoked Potential (SSVEP) based BCI system [81]. The SSVEP-based classifier was able to pick up the corresponding frequency from the EEG. The dataset used to train the generative models is the video-stimuli dataset [81] that contains 50 samples of EEG signals collected from offline videos played to one subject, referring to subject 1 in the NAO dataset [81]. The NAO dataset has two portions—offline and online—collected from tasks the same as in the Video-Stimuli dataset using a dry EEG device with 20 channels. The three generative models were trained only using the video-stimuli dataset, while the SSVEP classifier was tested on the NAO dataset. The generated EEG samples were used to pre-train the SSVEP classifier. The offline portion of the NAO dataset for each subject was fed into the pre-train model to fine tune for that particular subject. After the classifier was trained, the online portion of the NAO dataset was used to test the performance of the classifier. Different sizes of augmentation were empirically tested and compared. The result showed that for all three methods, a sample size of 500 resulted in the best classification accuracy. Table 3 shows the performances of different methods.

Table 3 shows all the papers used GAN as their DA technique. By reporting the magnification factor and accuracy before and after DA, we think that GAN technique is trending to use as DA technique for EEG signal.

Table 3. All reviewed papers that used GAN as their DA technique

Study	Dataset	Task information	Input formulation	Deep learning strategy	Best Performance Augmented Size	Accuracy (without DA)	Accuracy (with DA)
[61]	BCI competition II dataset III 1 subject 280trials 3 channels 2 class	Motor Imaginary EEG 7-15 Hz	Calculated features spectrogram	CNN cDCGAN	Doubled	83%	86%
[62]	30 subject 70000 trials 1 channel 2class	Mental task(EEG-Based Login Authentication)	Calculated feature Power spectral	GAN ,XGBoost	Added 10,000 samples	90.8%	95%
[63]	BCI competition IV dataset 2b 1subject, 400 trials, 3 channel 2 class	Motor imagery	Calculated features spectrogram	GAN+CNN	10 times	77.6%	79.3%
[65]	1 subject, 438 trials, 1 electrode 2 class	Motor task	Calculated features-spectrogram	GAN-SWD	NA	NA	NA
[69]	SEED 15 subjects, 62 channels, 3394 samples per subject	Emotion recognition 1-50 Hz	Calculated features spectrogram	CWGAN	Doubled	83.99% Arousal 69.02%, Valence 53.76%	86.96% Arousal 78.17%, Valence 73.89%
[69]	32 subjects, 32 channels, 2400 trials each subject	Emotion recognition 1-50 Hz	Calculated features spectrogram	CWGAN	Doubled	NA	NA
[72]	SEED: 9 subjects, 62 channels, 3classes, 45 videos/subject	Emotion recognition 1-50 Hz	Calculated features	cBEGAN	2000	81.9%	87.56%

	SEED V: 16 subjects, 62 channels, 5 classes		Differential entropy		1000	54.3%	62.8%
[73]	23 subjects 5085 trials more than 23 channels	Seizure	Raw signal	CNN	NA	81%	84%
[74]	18 subjects 32000 samples 14 channels	Emotion recognition	Raw signal	GAN	~triple	97.9%	98.4%
[75]	9 subjects, 32 channels, 500samples/subject	Motor imagery 0.5-100 Hz	Raw signal	CNN+LSTM	NA	NA	76.4%
[76]	10 subjects, 256 channels, 5 hours/subject	RSVP 0.1-55 Hz	Raw signal	CNN	3828	NA	NA
[81]	video stimuli dataset: 1 subject, 20 channels, 50 unique samples for each of the three class NAO dataset: 3 subjects, 20 channels 50 samples per class offline, 30 samples per class online	Steady state visual evoked 9-60 Hz	Raw signal	CNN	500	NA	NA

3.6.3. Sliding window or overlapping window

O'shea et al. (2017) presented a novel end-to-end architecture that learns representations from raw EEG signal by CNN for the task of neonatal seizure detection [82]. Interpretation of neonatal EEG requires highly trained healthcare professionals and it is limited to specialized units. They used overlapping window to augment 1389 seizures during 835 hours of EEG signal. Each trial split into 8s epochs with 50% overlapping to have more training sample for their proposed CNN.

They obtained 97.1% accuracy; however they didn't evaluate their result without overlapping or different shift lengths.

N. Kwak et al. (2017) used CNN for the robust classification of a steady-state visual evoked potentials paradigm [49]. They recorded EEG for the brain-controlled exoskeleton under ambulatory conditions. For generating more training samples, they used overlapping window. In their results, different shift lengths from 10 ms to 60 ms out of 2-s window were compared. They found the training samples with smaller shifts, performed much better than larger ones. The highest accuracy was 99.28% for 5-class visual evoked potential task.

For Schirrmeister et al. (2017), a key question was the impact of CNN training (e.g., training on entire trials or cropping within trials) on decoding accuracies [83]. The concept of overlapping window was pushed even further in this study: First, DA by overlapping windows share information was used to design an additional term to the cost function, which further regularizes the model by penalizing decisions that are not the same while being close in time. Second, redundant computations due to EEG samples being in more than one window were simplified, which ensured these computations were done once, thereby speeding up training. As a result, cropped training (segments of about 2 s length) increased the accuracy to 95% for CNN on high pass filtered data (The authors did not report the accuracies before DA).

Ullah et al. used a 1D-CNN for research on epilepsy detection [84]. The number of trials collected in this study was not enough to train the CNN. And obtaining a large-enough dataset during seizure activity was not practical. At the same time, the available, small dataset resulted in overfitting. To overcome this problem, the authors proposed 2 methods for DA:
(Note that the EEG signal length in this dataset was 4097):

- Sliding window of length 512, stride 64, leading to 87.5% overlap. Each of these windowed signals was treated as an independent instance. Therefore, each trial was divided to 57 sub-signals.
- Sliding window of length 512 with stride 128, leading to overlap of 75%, leading to 29 sub-signals

The average accuracies were 96.45±0.13 and 95.40±0.35 using DA with 87.5% and 75% overlap, respectively (The authors did not report the accuracies before DA).

N Truong et al. (2018) used GAN for semi-supervised seizure prediction [85]. They generated extra samples to balance the Freiburg and CHB-MIT datasets. As a result, training sets are 10 times larger than original one by using overlapping window. The extra generated training dataset is by sliding a 30-s window along the time with different shift length. However, they didn't report the accuracy achieved by different shifting length.

They achieved 60.91% and 72.63% accuracy (without DA) and 74.33%and 75.33% (with DA) for Freiburg hospital and CHB-MIT, respectively when training GAN on individual subjects.

Majidov et al. (2019) proposed an efficient classification of Motor imagery EEG task by using CNN [86]. For DA, they used sliding window with different shifting length. However, their result lacks more details about DA.

Z. Mousavi et al. (2019) proposed a single-channel EEG-based automatic sleep stage classification (2 to 6 classes) algorithm which processes the raw signals in order to learn features and automatically diagnose sleep stages using CNN [87]. The lack of balance between the data

of each class was challenging situation which caused biasedness of classification results and degraded accuracy. Therefore, they used overlapping technique to augment their dataset. The training set was 50% of the dataset included 7592 epochs (30s), however after DA, they had 24162 epochs (3s). They achieved to 93.55% accuracy for classification 6 classes of sleep stages. In addition, to evaluate the performance of the proposed DA, GAN was also implemented. However, according to their results, using GAN for the 6 sleep stages classification had achieved 72.33%, which is lower than overlapping window.

Avcu et al. (2019) developed an end-to-end CNN for seizure detection [88]. They strove to minimize the number of channels used (just 2 channels—Fp1 and Fp2) and compared that to the result with all channels. EEG data of 29 pediatric patients diognosed with a typical absence seizure were included in this study. In total, the data contained 1037 minutes of EEG with 25 minutes of seizure data distributed among 120 seizure onsets. To overcome the imbalance in the dataset, they applied different overlapping proportions according to existence or absence of seizures. Namely, while shifting with 5 seconds (no overlapping) was implemented to create interictal class, 0.075 second shifting was used for ictal class to create balanced input for the CNN. The sensitivity for 2-channel was 93.3% and for 18-channel was 95.8%. However, the result of DA was not reported in this study.

Tayeb at al. (2019) developed three deep-learning models: LSTM, CNN, and RNN for decoding motor imagery [89]. This group used shifting window with 4s length to reflect the partial time invariance of the data and overcome the problem of overfitting. This cropping strategy increased the training dataset by a factor of 25. The CNN architecture showed better performance and achieved a mean accuracy higher than 84% over all the 20 participants. However, their result lacks more details about DA.

Also, we found more papers which segmented the dataset to create more training data: Chambon et al. (2017) segmented the input data to 30s segment to create more dataset for each class of sleep stage [46]. Tsiouris et al. (2018) used LSTM for the prediction of epileptic seizure [90]. To overcome unbalance problem of rare seizure event, the EEG segment from the interictal class were split into smaller subgroups of equal size to the preictal class. Tang et al. (2017), proposed CNN for the failure prediction [91]. To avoid multiple instance learning issue for their CNN, they used segmentation window to have sufficient new training dataset. The length of each segment was found by adaptive Multi-scale sampling. Their result was improved from 70.9% (without DA) to 77.9% (with DA) on seizure dataset.

Although many studies used this method, there seems to be no consensus on the best overlapping percentage to use, e.g. the impact of using a sliding window with 10% overlap versus 90% overlap. Some studies tried different shifting length; but this issue still is not clear. For more information refer to Table 6.

3.6.4. Sampling

3.6.4.1. Oversampling

R. Manor et al. (2015), presented a CNN model for the use of single trial EEG classification in five category rapid serial visual tasks [92]. They used oversampling of the minor class (bootstrapping) to balance the dataset. They mentioned that although this method caused some overfitting on the minor class, however, it provided a more balanced classification performance in their experiment.

Drouin-Picaro et al. (2016), proposed a CNN model to classify saccades from frontal EEG signals to aim cursor control without the need for a separate eye tracking device in provide brain-computer interfaces [93]. In order to have a balanced dataset, horizontal saccades were sampled from without replacement so that the number of horizontal saccades in the dataset was the same as the highest number of vertical saccade (either up or down). The other vertical direction was then augmented by sampling from it with replacement, to make the number of data points in each direction equal. Hence, the dataset contained roughly 3000 examples of each saccade direction.

Supratak et al. (2017), used a CNN model, named DeepSleepNet, for automatic sleep stage scoring based on raw single-channel EEG. They extracted time-invariant features and used LSTM to learn transition rules among sleep stages automatically [94]. By duplicating the minority sleep stages in the original training set such that all sleep stages have the same number of samples they avoided overfitting.

Dong et al. (2017), proposed a Mixed NN for temporal sleep stage classification [95]. Because of the inherent imbalance in occurrence of the different sleep stages, the authors used oversampling to generate a new balance dataset which every sleep stage is equally presented.

Sors et al. (2018) used a CNN on raw single-channel EEG signal for scoring 5 class sleep stage [45]. They mentioned their dataset (SHHS) has a very imbalanced class distribution. In order to account for this, they tried cost-sensitive learning or oversampling but the overall performance using this approach did not improve.

Ruffini et al. (2019), randomly replicated subjects from the minority class to balance their classes [96]. Their proposed model helps for diagnosis derived from a few minutes of eye-close resting EEG signal collected at baseline idiopathic patients. They didn't compare the result with and without DA.

Sun et al. (2019) scored the sleep stage automatically. This study presents a stage-classification method based on a two-stage neural network [97]. The first, feature learning stage can fuse network-trained features with traditional hand-crafted features. A second, RNN stage is fully utilized for learning temporal information between sleep epochs and obtaining classification results. Oversampling was used to solve a serious sample imbalance problem. Sadly, the result lacked more details about DA.

3.6.4.2. Subsampling

Thodoroff et al. (2016), evaluated the capacity of a deep NN to learn robust features of EEG to automatically detect seizures [98]. They randomly subsampled the majority samples of the dataset to re-balance the ratio between seizure and non-seizure data (from 1000/1 to 80/20) which facilitate the training. However, because seizure manifestations on EEG are extremely variable both intra- and intra-patients, a second challenge was the overlack of data for each patient (average of 8 seizure per patient). They trained the CNN by using 0.5 s window instead on 1 s. Using transfer learning, the general representation of a seizure on other patients learned first and then they trained the model to the specific patient using the weights previously learned as initialization.

Sengur et al. (2019) employed deep feature extraction for focal EEG signals [99]. The deep features were extracted from spectrogram images using the AlexNet, VGG16, VGG19, and ResNet50 CNN models. The FC6 and FC7 activation layers were used for feature extraction resulting in 4096 dimensional feature vectors. The obtained feature vectors were used as input to various k-NN classification models. Random subsampling was performed as the DA technique (no other details were provided about the parameters). See Table 6 for more details about these studies.

Note. There are various resampling strategies to solve this type of problem such as random under sampling, random oversampling and random under sampling with synthetic minority over-sampling technique(SMOTE) [100-102]. We could find some studies which used SMOTE on EEG but they didn't satisfy our criteria [35, 103]. Beside resampling techniques, another way to deal with imbalanced data is cost sensitive learning [104]. Resampling strategies and cost sensitive learning can significantly improve the predictive evaluation of the models [105].

3.6.5. Fourier Transform

J. Schwabedal et al. (2018) proposed a new method for augmenting EEG signals when attempting sleep-stage classification [106]. They focused on imbalanced dataset in transitional sleep stages, such as S1 and S3, which are rare events with respect to more stable stages such as wakefulness or Rapid Eye Movement (REM) sleep. Cost-sensitive learning [45], oversampling of the minority class [92, 94, 107], and subsampling the majority class [98, 108] are common techniques to address imbalanced classes. But the overall performance using these approaches resulted in some biases in prediction and did not improve the accuracy [1]. Therefore, they used Fourier Transform Surrogates to augment the EEG data. The complex Fourier components of a signal x_n can be decomposed into amplitudes a_n and phases φ_n:

$$x_n = a_n e^{i\varphi_n}$$

Under the assumptions of linearity and stationarity of the signal, they generated a new signal which is statistically independent from the original signal. This happened by randomizing the Fourier-transform phases [0, 2π] and then applying the inverse Fourier transform. The authors processed the CAPSLPDB sleep database, consisting of 101 overnight Polysomnographies (PSGs), using a CNN for 6 sleep stage classification. They then used the above method to

balance and augment the database to achieve better generalization. They improved the mean F1-score by 7% for sleep-stage classification.

Zhang et al. (2019), proposed a novel DL approach with DA to improve classification of motor imagery EEG signals [42]. They applied the empirical mode decomposition on the EEG frames and mixed their intrinsic mode functions to create new artificial EEG frames, followed by transforming all EEG signals into tensors as input for the NN by complex Morlet wavelet s. Complex Morlet wavelets transformation of the EEG signals has been proved effective in recent motor imagery researches, including tensor decomposition and wavelet-based combined feature vectors method. Their algorithm decomposes the original signals into a finite number of functions called intrinsic mode functions (IMFs). Each of these IMFs, represents a non-linear oscillation of the signal. Once the signal has been decomposed, it can be recovered by adding all IMFs and the residue without loss. The main idea in this study is that by mixing IMFs of the same class we can generate new samples from this class by preserving all intrinsic characteristics. This aims to decrease overfitting problem in training NN and eventually improves classification results. They used CNN and WNN (Wavelet Neural Networks) models to evaluate their results.

They found that magnifying two times the original training sets had highest mean value and better stability in CNN. And as for the WNN, the highest magnification was achieved by 5. The average of the accuracy for CNN was better than WNN. They evaluated their method on BCI competition dataset. By magnification factor 5, the CNN accuracy was 77.9% without DA and 82.9% with DA and WNN reached 88% without DA and 84.3% with DA. The relatively low computational efficiency of the WNN was the limitation in their proposed work. This group worked very well on DA details. They used two more big motor imagery datasets to evaluate their methods. They found that WNN has better classification performance and smaller loss than the CNN. However, each iteration of the WNN model takes almost five times as long as the CNN. And they speculated that it's because the WNN lacks the consideration of parallel computing.

3.6.6. Recombination of Segmentation

Said et al. (2017) presented a joint compression and classification method for EEG and electromyogram (EMG) using a multimodal auto encoder [109]. They conducted their experiments on the DEAP dataset. It included the modalities of EEG, EMG, and multiple physiological signals recorded from 32 participants during 63 seconds at 128 Hz. During experiments, volunteers watched 40 music videos and rated them on a scale of 1 to 9 with respect to four criteria: likeness (dislike, like), valence (unpleasant to pleasant), arousal (uninterested or bored to excited) and dominance (helpless and weak feelings to empowered feelings). Signals were normalized and segmented into 6 seconds segments. EEG and EMG modalities contained 23040 samples of 896 features. They trained the multimodal auto encoder by adding zero values to one modality while keeping the original values for the other modality and vice-versa. Thus, one third of the training data was EEG only, another one third was EMG only, and the rest had both EEG and EMG data.

Zhang et al. (2019) used common spatial pattern (CSP) and CNN to detect seizures [110]. They first split each training EEG trial into three segments, and then generate new artificial trials as a

combination of segments coming from various, randomly selected trials. They achieved 90% average accuracy, but did not report their multiplication factor or the accuracy before DA.

Dai et al. (2019) employed hybrid scaling CNN (HS-CNN) for motor imagery classification [111]. They varied the CNN kernel size between subjects and even between sessions. They found three kernel sizes for each selected frequency band (theta, mu, and beta). To improve the accuracy of HS-CNN, they used a 3-stage DA method: (1) segment each trial to 3 segments; (2) recombine the segments within different trials in the time domain; (3) swap frequencies: after band-pass filtering, the filtered trials (theta, mu, and beta) in the same frequency band were randomly swapped. Step2 and 3 were repeated multiple times for a multiplication factor of 3. The average accuracy for dataset 2b of BCI competition IV increased from 86% to 87.6%. They tried other DA techniques, such as noise addition and sliding window resulting in average accuracies of 86.1% and 80.1%, respectively.

3.6.7. Other

Frydenlund et al. (2015) used video and EEG data from subjects to estimate emotional response to music video (120 one minute music videos) [112]. To reduce computational cost, the researchers often throw away part of the signal by downsampling. In this experiment, authors reused the data thrown away during downsampling as new trials. Downsampling by a factor of N would therefore allow an augmentation of N times. However, the authors did not explicitly frame this as a DA method. So, no direct comparison was made of the accuracy with and without using the downsampled data.

Sakai et al. (2017), published a paper about DA methods for ML-based classification of bio-signals [48]. Their proposed DA methods for EEG signals includes: a) Shifting all-time data ($\pm$10ms) b) Amplifying all-time data (90% and 110%) c) Shifting near-peak value ($\pm$10ms) d) Amplifying near peak value (90% and 110%). Multiplication factors ranged from ($\pm$5% to $\pm$50% every $\pm$5% in b and d and $\pm$5ms to $\pm$5ms every 5ms).

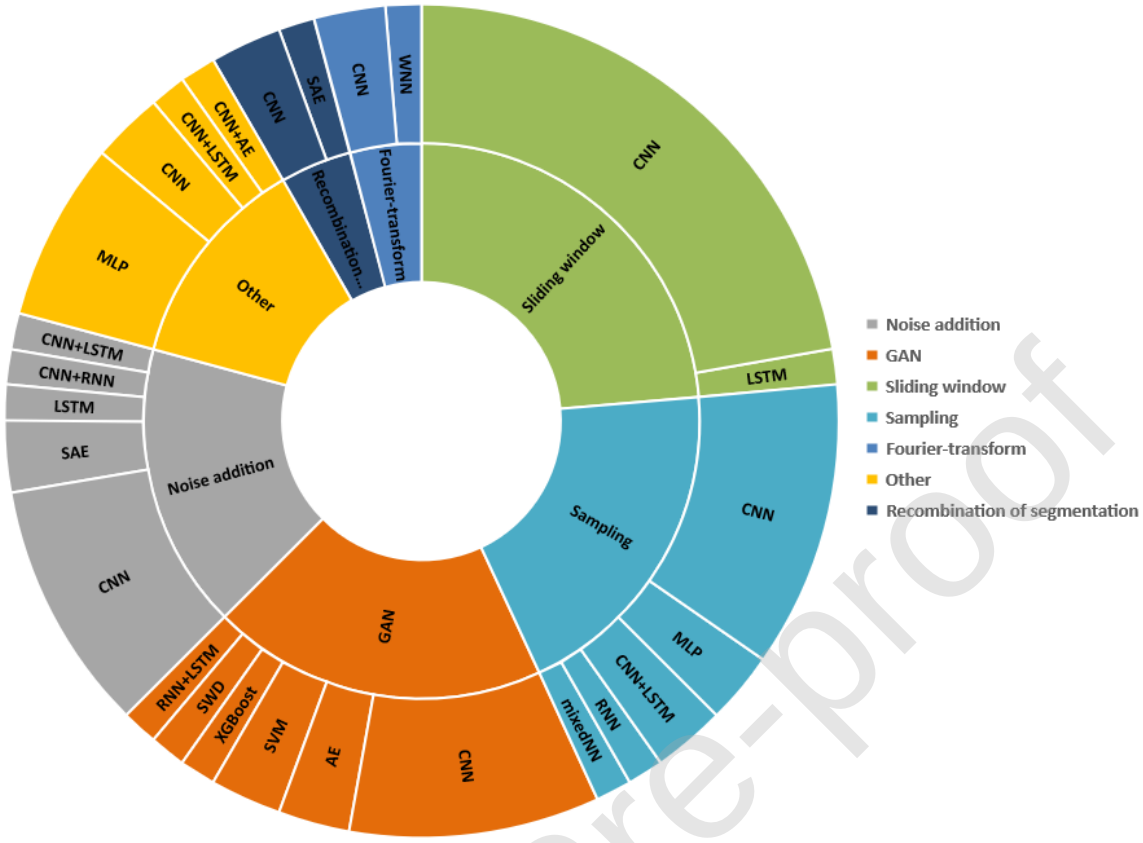

Figure 10. Data augmentation methods across all reviewed papers. The inner circle shows the general DA methods and the outer circle shows the deep learning architecture strategy used.

Deiss et al. (2018) suggested swapping right and left electrodes to double the size of the dataset. They utilized a dataset of brain monitoring in an intensive care unit (ICU) for 5-way classification (Seizure, Lateralized Periodic Discharges (LPD), Generalized Periodic Discharges (GPD), Generalized Rhythmic Delta Activity (GRDA), Lateralized Rhythmic Delta Activity (LRDA)), and the last one corresponds to Other/Artifacts (O/A))on 155 patients [113]. The most challenging issue in their experiment was to make the model learn how to generalize to new patients. To simulate different patients, they kept three reference electrodes in the middle of the scalp unchanged and left/right flipped the remaining electrodes. Swapping electrodes in this manner doubles the amount of data. The authors reported that this DA method did not affect classification for tasks with symmetrical signals between the brain hemispheres (The authors did not report the accuracies before DA).

Shovon et al. (2019) applied STFT on EEG signals to transform signal to images for binary classification of motor-imagery signals [114]. They used rotation, flipping, zoom in and zoom out as DA techniques to overcome the overfitting problem in their proposed CNN model. Additional 1000 augmented images increased the average accuracy to 89.19% (no accuracy before DA was reported).

Freer et al. (2019) constructed a convolutional LSTM (C-LSTM) network based on filter bank common spatial patterns (FBCSP) for 4-way classification in a motor-imagery task [115]. The

effects of several DA methods of data augmentation on different classifiers were explored, combining noise addition, multiplication, frequency shift, and phase shift. These DA methods improved the average overall accuracy of the classifiers by 5.3%.

Finally, Mokatren et al. (2019) applied the discrete-wavelet transform to extract energy and entropy of 4 frequency bands: theta, alpha, beta, and gamma in an emotion-recognition task [116]. A 3-D array of size KxKxB was created, where the first two dimensions represent an image of KxK pixels corresponding with the channels positioning over the scalp, while the third dimension represents the number of features: energy and entropy for 4 frequency bands(B=8). They used image augmentation techniques, such as horizontal and vertical shifting, to improve the accuracy of their CNN. Their classification accuracy on the DEAP dataset improved from 86.47% (without DA) to 90.87% (with DA) for Arousal and 88.34% (without DA) to 91.33% (with DA) for valence.

This section shows that these authors tried to improve the accuracy of their classification method with different techniques but because we found just one case from each of this innovation, we grouped them together. Figure 10 displays the aggregated information on DA methods and DL architecture strategy. While there is no clear consensus when looking to all 53 studies together, studies that employed sliding window, sampling, and noise addition as DA method, mostly used CNN. We investigated the EEG task compared across different DA techniques Figure 11. Following our review, we conclude that, for seizure task, the sliding window method should be used. For Mental workload, noise addition achieved the best results. And for deciphering sleep stages, the sampling method is the best fit. In sum, we recommend that sliding windows should be used for seizure detection. We also found that noise addition works best for mental workload. And the sampling method appears optimal to classify sleep stages.

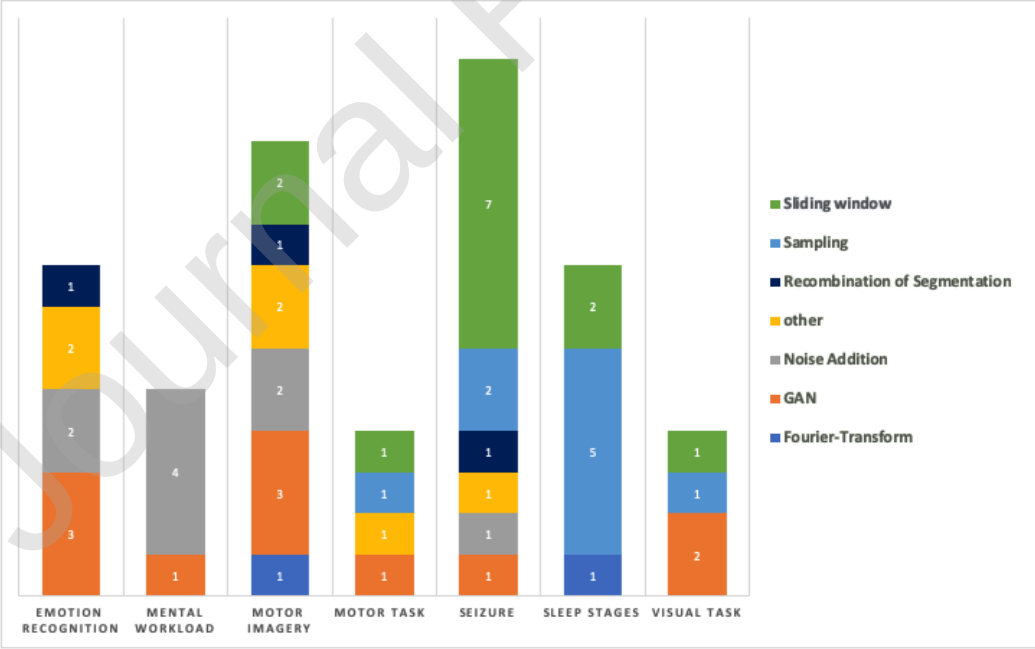

Figure 11. Number of papers for general EEG tasks compared across different DA techniques

3.7 Accuracy gains of data augmentation

The application of DA for DL on EEG is still nascent, with relatively few studies having been conducted. What is more, many of those studies unfortunately do not report the gain in accuracy that the DA method brought about, or which parameters were used exactly (Table 6). Nevertheless, 29 of the 53 papers we surveyed included a measure of accuracy before and after DA. We therefore computed an improvement score on those for each DA analysis, $\frac{a-o}{1-o}$. Here, "a" stands for the accuracy of the model when trained on the augmented dataset and "o" stands for the accuracy on the initial, non-augmented dataset. Hence, an improvement score of s suggests that, by training also on the augmented dataset, a fraction s of the gap between initial accuracy and perfect accuracy was covered by the model trained on the augmented dataset. The overall improvement score was 0.29±0.08 (mean ± s.e.m.). Though the score varied among the different DA techniques—from 0.08 for recombination of segmentation to 0.36 for noise addition (Fig. 12A). For tasks, it varied from 0.14 for motor imagery to 0.56 for mental workload (Fig. 12B). The 95% confidence intervals for all tasks (except "visual task") and DA techniques did not include 0. The above therefore suggests that various DA techniques for EEG signals of different tasks does improve the classification accuracy for various augmentation methods. It should be noted though that these statistics rely on relatively small number of analyses. And thus, more studies are required to establish reliable DA improvement score for different techniques and tasks.

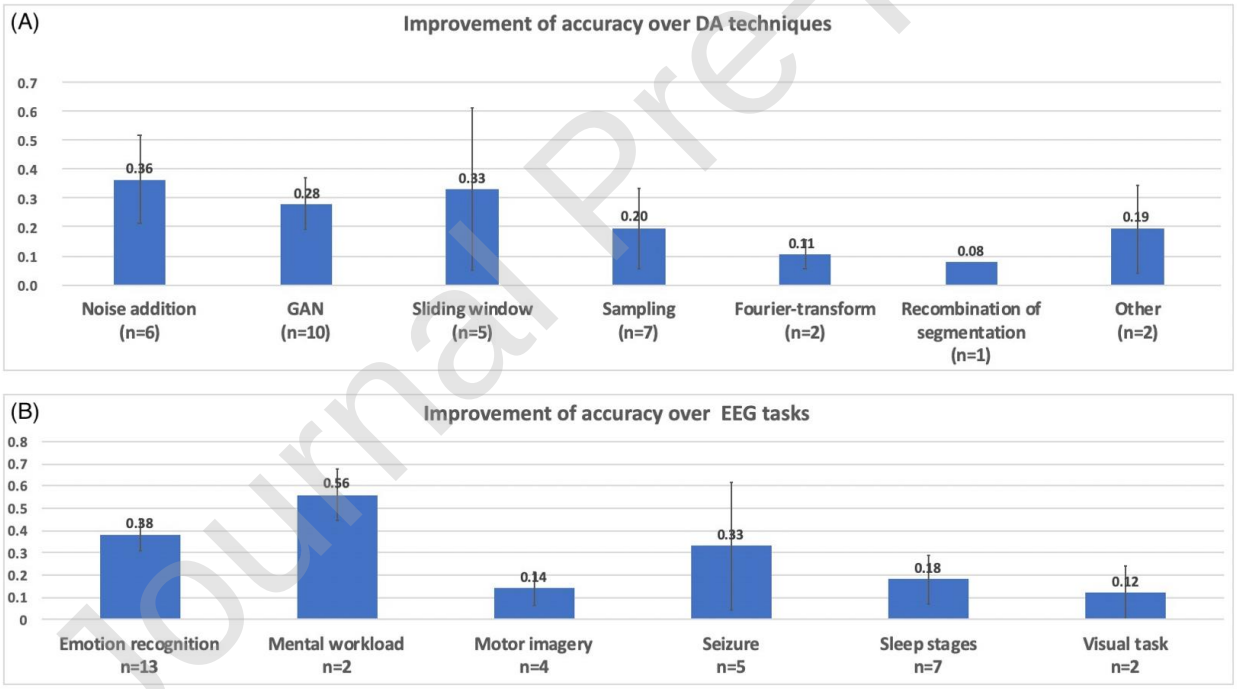

Figure 12. (A) The improvement score, or the fraction of variance left unexplained by the original DL method that was explained when training the model using DA, for different DA techniques (mean ± 95% confidence intervals). Here'n' is the number of studies everywhere except GAN, Sampling, and Other, where there were more than one analysis, with different accuracies, reported in each study; hence there 'n' is the number of accuracies. (B) Same as A but the improvement is over EEG tasks. Here "n" is the number of studies everywhere except "emotion recognition", where there were 9 studies, 2 of which ran multiple DA

analyses; hence "n" there is the number of analyses. No motor-task studies included accuracy before and after DA, so that task is not included in this figure.

3.8 Data augmentation for EEG—a working example

The literature we reviewed suggests that various DA techniques improve classification accuracy for various EEG tasks. But we nevertheless wanted to test this out for ourselves. Therefore, in this section, we show one example where we compare all the data augmentation techniques reviewed above on a specific EEG task based on our analyses of those techniques and methods. Figure 2A suggests that motor imagery was the most popular EEG task for DA in 2019. This led us to select the 2008 BCI competition IV dataset 2a as our motor imagery example [111, 117-119]. We selected left and right, as this was the most common classification technique in the literature and thus facilitated the comparison of our result with the literature [111]. Based on Figure 4B, raw signals were at least as commonly used as other features. As these were also the least processed type of signal, we selected these raw signals as inputs for our NN. Following Figs. 5-7, and 10, we concluded that the CNN architecture is likely to be the best fit for our data. Further based on Figure 6, which aggregated information for deep learning architectures, we used a CNN with conv(2) + FC(2) together with a leakyRelu activation function. From Figure 11 we found that we should test all the DA techniques. In addition, we evaluated the result without using DA and also with different magnification factors.

We ran various DA techniques with different magnification factors on the dataset (Table 4). We found that GAN with a magnification factor of 15 had the best accuracy overall for this dataset. We further compared the improvement score of all the DA techniques, each for its optimal magnification factor (Fig. 13). It is clear that GAN improves the classification accuracy more than all other DA techniques.

Looking through the literature, the best accuracy we could find on the BCI competition IV dataset 2a dataset was for from a study by Dai and colleagues [111]. We therefore compared the accuracy for all 9 subjects in the dataset between that paper, our analysis without data augmentation, and our analysis with data augmentation (Table 5). Our accuracy was higher than Dai and colleagues by more than 2%, with an improvement score of 0.24. What is more, data augmentation improved our results (over the same technique without data augmentation) by more than 4%, resulting in an improvement score of 0.41.

Table 4. Different data augmentation techniques and magnification factors that we used on the BCI competition IV dataset 2a dataset. Mean results over all subjects.

DA techniques		**Fourier-Transform**	**Noise Addition**			**GAN**		**Sliding window**
parameter for each DA		**(EMD)**	**σ =0.1**	**σ =0.2**	**σ =0.5**	**Conditional (left vs. right)**	**Conditional (left vs. right and channels)**	**sliding window of length 125**
Magnif	2	0.8671	0.9056	0.8982	0.8768	0.9133	0.9025	0.8948

	5	0.8652	0.8999	0.8849	0.8908	0.924	0.9092	0.8904
	10	0.8822	0.8902	0.892	0.8721	0.9087	0.9217	0.8992
	15	0.8858	0.8988	0.8756	0.875	0.9358	**0.936**	0.8949
	20	0.8932	0.8898	0.8975	0.8904	0.9193	0.93	0.9092

Table 5. Comparison of accuracies on BCI competition IV dataset 2a between the best result we could find in the literature, our best analysis without DA, and our best analysis with DA

subject's ID	Recombination of segmentation [111]	Our proposed method/without DA	Our proposed method with the best DA (GAN with m=15)
1	90.07%	91.58%	95.38%
2	80.28%	89.67%	91.25%
3	97.08%	91.89%	91.25%
4	89.66%	90.05%	96.12%
5	97.04%	91.28%	95.05%
6	87.04%	90.97%	94.62%
7	92.14%	81.38%	91.22%
8	98.51%	91.20%	90.54%
9	92.31%	83.95%	97.50%
Average ($\pm SE$)	91.57($\pm$1.9)%	89.11 ($\pm$1.2)%	93.60($\pm$0.8)%

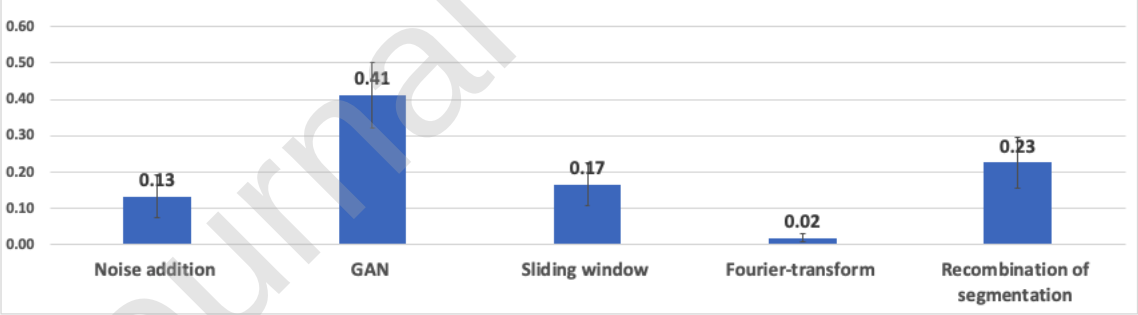

Figure 13: Improvement of accuracy over DA tehcniques on BCI IV-2a dataset (mean ± 95% CI).

4. Discussion

Here we review the most important findings from our results section and discuss the significance and impact of various trends highlighted in the results. We also provide some recommendations for the 7 tasks on which we analyzed DL_EEG: seizure detection, sleep stages, motor imagery, mental workload, emotion recognition, motor tasks, and visual tasks.

4.1. Rationale

The relatively small size of EEG datasets drastically decreases the effectiveness of DL. In the past few years, DA techniques have received widespread attention and achieved considerable performance gains for DL. Therefore, we focused our review on available DA methods for DL-based EEG. Such augmented datasets facilitate training more complex models, with more parameters, while at the same time potentially reducing overfitting. We only considered papers that focused on DA in DL-based analysis of EEG.

Previous review papers recommended that more targeted work be carried out to fully exploit the potential advantages of DL in EEG processing **[1, 23]**. It thus appears natural to explore the relation between performance and DA. Toward this goal, we carried out a systematic review of DA for DL-based EEG. Our goal was to address the following critical questions: (1) What DA approaches exist for EEG? (2) Which datasets and EEG classification tasks have been explored with DA? (3) Are specific DA approaches more suitable for particular tasks? (4) What input features are used for training deep networks with DA?

4.2. Data

A lingering critical question in machine learning is "how much data is enough data?", and it is of special relevance when applying sophisticated DL techniques on limited size EEG datasets. Naturally, the amount of data is critical in achieving high DL performance. But, needless to say, the quality of the data is also very important. To analyze this, we looked at dataset features such as the number of subjects, amount of EEG recorded (in trials or time), and the DA schemes used. We found that, for noise addition, the best results were obtained when a lower standard deviation was used. However, more generally, we did not find one specific, definitive answer to the data quantity question. That said, our analysis clearly suggests that DA techniques are typically successfully able to increase the performance of DL (Table 6).

4.3. EEG preprocessing

Most studies used frequency-domain filters to limit the bandwidth of the EEG signals. This enabled them to focus on specific frequency ranges that were of interest (Figure 3). The filtered frequency ranges were organized by EEG task type. We found no studies that specifically tested the role of this filtering on NN. (This lacuna is discussed in other review papers **[23]**.) The great majority of the reviewed papers preprocessed the EEG data before feeding it into NNs. Based on Figure 4, 49% of the reviewed papers used calculated features such as wavelet, entropy, spatial filter, or STFT as the input to NNs. On top of that, 36% simply used the raw EEG time-series signal as the only input to the NN. This is not surprising as a key motivation for using NN for EEG processing is to automatically learn features. An analysis of the sort that we carried out could in principle give some sense of which input types should be used for these purposes. But a complete answer depends on many factors, including the EEG task. And it is therefore difficult to draw definitive conclusions when only 53 studies using DL and DA are currently available.

4.4. Deep-learning methodology

Our analysis focused on architecture trends and input formulations for each architecture. However, the EEG task too is of importance, of course. Based on Figure 6, CNN was the most popular NN architecture—likely because it is well suited for end-to-end learning, scales well to

large datasets, and can exploit hierarchical structure in natural signals. The number of hidden layers in the different NN architectures varied case by case. Given the relatively small number of papers so far, we were able to be aggregated information about DL architectures into a single figure (Fig. 7), which we hope would help our colleagues gain some intuition about this nontrivial issue. Thus Figure 7 shows that the input formulation for RNN is images while for SAE it is calculated features. For LSTM, there are two input categories: signal and calculated features. CNN and MLP studies included instances of all input formulations but the signal formulation was used most often for their inputs. There were also hybrid architectures that used a combination of two standard NN. The papers relying on such hybrid NNs commonly used calculated features and images as their inputs.

4.5. Data augmentation

Figure 8 is a testament to the importance of DA for EEG processing with DL. DA techniques have received widespread attention and achieved appreciable performance boosts for DL techniques on EEG. However, more work is required to clearly assess their advantages as well as their potential disadvantages. Here we covered all the available DA techniques that we could systematically source and grouped them into 7 categories: noise addition, GAN, sliding window, sampling, Fourier transform, recombination of segmentation, and other. Sliding windows, at 24%, was the most common. Nevertheless, there seems to be no consensus on the best overlapping percentage to use between consecutive windows—e.g., the impact of using a sliding window with 10% versus 90% overlap. Some studies tried different shifting lengths **[49]** [84] , but this issue remains unsettled.

We found two main approaches for adding noise to EEG signals for DA: (1) adding various types of noise (Gaussian, Poisson, salt and pepper, etc.) to the raw signal; (2) converting the EEG signals to sequences of images (spectrograms) and adding noise to these images. Though it has been reported that adding noise to the images did not improve the classification accuracy **[50]**. Unfortunately, some authors did not provide details about the accuracy before and after DA. In a similar vein, critical noise parameters (e.g., mean, standard deviation, the magnification factor of training dataset) were sometimes not reported. This made it more difficult to compare techniques and parameters across studies.

Since 2018, GAN has become very popular for generating EEG signals that mimic real ones. Though GAN and related DL algorithms were used and discussed more for generating synthetic images for image classification tasks. EEG can often be analyzed and visualized in the frequency domain over time as spectrograms (through a Fourier or wavelet transformation). These spectrograms can then be treated like any other image, and therefore data augmentation methods that were developed for images can, at least in the technical sense, be directly applied to them. The spectrograms generated via the DA process is then converted back to an EEG signal of course.

While GAN data augmentation for EEG shows some improvement of classification accuracy, it has still not been clearly demonstrated to be better than other, simpler methods—like noise addition. For example, from our results (Table 6), it appears that the mean increase in accuracy when using GANs is 5.7% (STD 5%) while for noise addition, the increase is 14.2% (STD 13%). The GANs that we covered here mostly learned the underlying distribution of the training EEG

signals and generated fake signals that are within the distribution. In this sense, it might be argued that it is not that different from adding noise to the original signals. In addition, whether we can treat a spectrogram as an image and simply apply image-based data augmentation techniques remains an open question. First, the "pixels" in spectrograms relay temporal and frequency information that images do not. Second, at least when using CNNs, the invariant filters that work well across images would often not be expected to work on the spectrograms. For example, there is no a priori reason to expect to find the same pattern in high gamma early in time and in theta at a later time. Third, when using real images, the developer of a GAN can rely on her visual system to judge how well the GAN works for generating fake images that are similar to the real ones on which it is based. However, the same cannot be said for a GAN developed for EEG. What is more, we know too little about the characteristic of EEG for specific tasks (certainly across subjects and variants of the task) to develop a method that would judge the quality of an EEG GAN. So, this technique should be used with proper caution.

Fourier transform was used in 2018 to augment EEG signals, very successfully. This method assumes linearity and stationarity of the EEG signals [106] . In 2019, Zhang and colleagues used these intrinsic features of EEG and decomposed the signal to its IMFs. By mixing IMFs, they generated new samples and decreased overfitting [42]. Sampling was used in many studies to better balance imbalanced datasets. Balancing the number of samples among classes may drastically improve the usefulness of a dataset.

Figure 11 enables us to draw a few trends. We see that the sliding-window technique is used for the majority of papers that analyze seizure detection. Similarly, we found that noise addition is the most common technique for mental workload. And sampling methods appear most often for classifying sleep stages. To what extent the more popular techniques are also more optimal is still unclear given the relatively small number of papers so far.

But how useful is DA for DL-based EEG analysis? How much does it improve classifier accuracy? Before delving into the numbers, it is also worth bearing in mind that publication bias might be driving these numbers up. Unsuccessful attempts at improving DL accuracy with DA may be less likely to be published. That said, on average, training on the DA dataset helped make up almost 3/10 of the gap that was left in accuracy between the original analysis and perfect accuracy. Though this improvement score varied widely among DA techniques and tasks (Fig. 12). Too few studies reported both accuracy before and after DA and the parameters of their DA method for us to be able to carry out more in-depth statistical analyses. Those will be possible with additional publications.

Given the above, and to further test the usefulness of DA for EEG, we ran our own EEG analysis of an open dataset (BCI IV-2a), based on what we learned from our review of the literature. We then compared the accuracy were able to obtain with DA to the best accuracy we could find in the literature on the same dataset (Dai and colleagues [111]) as well as to our technique without DA. We were able to explain 24% of the accuracy beyond that explained by Dai et al., which suggests that our analysis of the literature is useful in gaining insight into which DA techniques work well for EEG analysis. What is more, our DA-based EEG analysis was able to explain 41% of the accuracy that remained unexplained by the model without DA. These results therefore

increased our confidence that data augmentation for EEG, even though still nascent, is a promising venue for improving the classification accuracy of DL-based techniques for EEG.

4.6. Guidelines for reporting results in papers

Some papers clearly explained their methodology with respect to DA (e.g., [115]). Unfortunately, these were the exception. Of the reviewed papers, 45% did not report the accuracy before DA and 38% did not report the parameters they modified in their DA method. It is also noteworthy that 41% did not mention the magnification factor they used. This made surveying and comparing the literature rather difficult.

Therefore, in order to improve the quality and reproducibility of the work in the field of DA on DL-based EEG, we recommend that authors follow the guidelines below when reporting their results in their studies.

• Clearly describe the data augmentation method they used	Method, parameters they change in their method, Magnification factor
• Clearly describe the dataset	#subject, #trials, #classes
• Test their proposed method on an existing dataset	Compare model performance and evaluate their results on a public dataset
• Clearly describe the architecture	#layers, their widths, the activation functions used
• Report the accuracy and results	Report the accuracy before and after using DA, report the results when changing the parameters of DA
• Share internal recording and reproducible code	Whenever possible, including hyperparameter choices

5. Limitations

One clear limitation of our study is the relatively small number of papers published so far on this topic. This precludes us from carrying out mode detailed analyses than the above. What is more, another obvious limitation of our methodology, already discussed above, is that our analysis is only as good as the data on which it is founded. When little information is provided about the DL or DA methods, it directly and immediately limits our ability to analyze those data, as discussed above.

In addition, although the search methodology we used to identify relevant studies is well-founded, it undeniably did not capture all of the existing literature on the topic. Since the field of DA for DL-based EEG is still young and the number of publications available at the time of writing this manuscript was limited, we decided to include all the papers we could find (note that some of the newer trends are more visible in repositories such as arXiv and bioRxiv, as those manuscripts may be going through the publication process). They have been adopted by the DL community to quickly disseminate results and encourage a fast research-iteration cycle. Our goal was to provide a transparent and objective analysis of the trends in DA for DL-based EEG.

Also, in preclinical, experimental research and highly competitive environment for funding, researchers submit predominantly positive results for publication. Withholding negative results from publication — publication bias — could have major change in our analysis.

We focused our analysis on the points that we thought would be most interesting, valuable, and impactful for the performance of DA on DL-based EEG. Therefore, we didn't include normalization procedures, software toolboxes, loss function, training time etc., in this analysis.

6. Conclusions

DL has been successfully applied to many EEG tasks such as: sleep stages, motor imagery, mental workload, and emotion recognition tasks. Applying DL to EEG has shown great promise in processing these complex signals due to its capacity to learn good feature representations from raw data through successive non-linear transformations. However, DL is inherently limited over EEG datasets because of their relative smaller size. DA, in turn, increases the available training data, facilitating the use of more complex DL models. It can also reduce overfitting and increase the accuracy and stability of the classifiers

Looking at the inputs to the DL architectures, the most common technique is still to calculate features (49%) outside the NN and feed it into the network, though a sizable fraction of papers input the raw signals (36%) into the NN and let it extract features itself. In addition, while various architectures have been used successfully on EEG datasets, CNN is most often used (62%). Taking all of the above into account, our analysis of the literature suggests that DA was mainly used for seizure detection (24%) and motor imagery (21%). In particular, sliding windows are favored for seizure detection. Noise addition is most common for mental workload. And sampling methods are the procedures of choice to classify sleep stages.

Our attempt to compare results between different studies highlighted for us the high degree of variability in how results were reported across studies. We therefore added our own analysis of an open dataset that provided clear evidence that DA leads to a gain in accuracy for DL-based analysis of EEG. We further made specific recommendations to ensure reproducibility and better comparison of the results when the authors use DA and DL. It is key to clearly describe the DA method, its parameters and their role in achieving the accuracy that the paper boasts. It is also critical to report the magnification factor as well as the accuracy before and after DA.

In sum, we hope this review will constitute a good entry point for EEG researcher looking to apply DA for training DL algorithms on their datasets and will assist the field to produce high-quality, reproducible results.

6. Acknowledgements

This publication was made possible in part through the support of a joint grant from the John Templeton Foundation and the Fetzer Institute, and Boston Scientific Corporation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation or the Fetzer Institute. This publication was also made possible in part by the support of Boston Scientific Investigator-Sponsored Research Program.

Table 6. Details of all the papers that we found for our review paper. In Data Augmentation column: NA: Noise addition, SW: Sliding window, S: Sampling, FT: Fourier-transform, Recombination of Segmentation: RS, O: Others and in EEG task column: ER: Emotion recognition, MW: Mental workload, MI: Motor imagery, S: Seizure, SS: Sleep stages, IT: Imagery task, VT: Visual task and in input formulation column: S: signal, I: Images and CF: Calculated features. Some studies used different dataset or different DA techniques and we show them separately.

Study	Data Augmentation	EEG task	# Subjects	# Trials	# Channels	# Classes	Frequency Range	Input Formulation	Type (input formulation)	NN Architecture	Hidden Layers	Last Layer (# Classes)	Activation Function in Hidden Layers	Activation function in Output Layer	Accuracy (before DA)	Accuracy (after DA)	Dataset
[50]	NA	MW	13	2670 trials	64		4-13 HZ	I	FFT	CNN +LSTM	conv(7)	FC(512)	Relu	softmax	NA	NA	University of Memphis Institutional review board
[51]	NA	MW	8	180 min/subject	1	2	1.5-40 HZ	CF	power spectral	SAE	hid(5)	FC(2)	Relu	softmax	NA	0.93	SDAE
[52]	NA	MW	7	1h	11		1-40 HZ	CF	FFT, power spectral	SAE	hid(6)	FC(2)	NA	NA	0.342	0.75	AUTOCAM
[28]	NA	ER	14	1890 trials	62	3	1-50 HZ	CF	Entropy	CNN	conv(4)	FC(3)	Relu	softmax	0.49	0.74	SEED
[28]	NA	ER	14	1890 trials	62	3	1-50 HZ	CF	Entropy	CNN	conv(13)	FC(3)	sigmoid	NA	0.765	0.855	SEED
[28]	NA	ER	30	527 trials	32	3	1-50 HZ	CF	Entropy	CNN	conv(13)	FC(3)	sigmoid	sigmoid	0.408	0.45	MAHNOB-HCI
[53]	NA	S	5	16.3 h	100	2 & 3 & 5	0.53-40	S	Raw	LSTM	LSTM(1)	Softmax	ELU	softmax	subject by subject	increased	Bonn University
[47]	NA	MW	22	6490 trials(8h)	64	4	4-30HZ	CF	power spectral	RNN +CNN	hybrid= conv(9) +LSTM (1)	FC(4)	Relu	softmax	NA	0.93	NIMHANS
[54]	NA	ER	32	40(video 1 min)/subject	32	2 & 4	[1,50]U[60,end]	CF	spatio-temporal	CNN	conv(2)	softmax	NA	sigmoid	0.79 0.79	0.88 0.87	DEAP
[55]	NA	MI	5	5 sessions	3	2	0.5-100 Hz	CF	Raw	CNN	Conv(4)	FC(2)	ELU	softmax	NA	NA	BCI competition IV 2b
[56]	NA	MI	9	72 trials/task	22	4	7-125HZ	CF	spatio-temporal	CNN	conv(1)	FC(4)	Relu	sigmoid	0.709	0.779	BCI competition IV 2a
[56]	NA	MI	14	880 trials(160 trials include	128	4	7-125HZ	CF	spatio-temporal	CNN	conv(1)	FC(4)	Relu	softmax	NA	NA	High Gamma dataset(HGD)

				for test and train)												
[62]	GAN	MW	30	70000 trials/subject	1	2	NA	CF	power spectral	XGBoost	Unspecified	Unspecified	logestic sigmoid	softmax	0.90	0.95	NA
[61]	GAN	MI	1	280	3	2	7-15HZ	I	spectrogram	CNN	Unspecified	Unspecified	Relu	softmax	0.83	0.86	BCI compititionII dataset III
[65]	GAN	MT	1	438	1	2	8-13HZ	I	spectrogram	wasserstein distance(SWD)	Unspecified	Unspecified	leaky Relu	softmax	NA	Did not improved results	not mentioned
[69]	GAN	ER	15	3394 per subject	62	3	1-50HZ	I	spectrogram	CWGAN	Unspecified	Unspecified	Relu	softmax	0.839% Arousal 0.69%, Valence 0.53%	0.869% Arousal 0.78%, Valence 0.73%	SEED
[69]	GAN	ER	32	2400 per subject	32	2	1-50HZ	I	spectrogram	CWGAN	Unspecified	Unspecified	softmax	softmax	NA	NA	DEAP
[63]	GAN	MI	1	400trials	3	2	NA	I	spectrogram	CNN	Unspecified	Unspecified	NA	softmax	0.77	0.79	BCI cometitionIV dataset 2b
[57]	GAN	ER	9	45video/subject	62	3	1-50Hz	CF	Entropy	cBEGAN	Unspecified	Unspecified	Relu	NA	0.81	.87	SEED
[57]	GAN	ER	16	48 experiment	62	5	1-50Hz	CF	Entropy	cBEGAN	Unspecified	Unspecified	Relu	NA	0.54	0.62	SEED V
[73]	GAN	S	23	5085 trials	>23	2	NA	S	Raw	CNN	Conv(5)	FC(2)	Relu	softmax	0.81	0.84	CHB-MIT
[74]	GAN	ER	18	32000 trials	14	2	NA	S	Raw	GAN	NA	NA	NA	NA	0.97	0.98	Not public

[75]	GAN	MI	9	500trials/subject	32	2	0.5-100Hz	S	Raw	GAN	CNN+LSTM	Unspecified	Relu	softmax	NA	0.76	BCI cometitionIV dataset 2b
[76]	GAN	VT	10	5 hours/subject	256	2	0.1-55Hz	S	Raw	CNN	conv(2)	Unspecified	NA	NA	0.50 (cross subject) 0.62(same subject)	0.53 (cross subject) 0.627(same subject)	BCIT X2 rapid series visual presentation
[81]	GAN	VT	1	50trials/class	20	3	9-60Hz	S	Raw	CNN	conv(1)	FC(3)	Relu	softmax	NA	NA	Video-Stimuli Dataset for augmentation, NAO Dataset for testing
[81]	GAN	VT	3	50 offline trials/class+ 30 online trials/class	20	3	9-60Hz	S	Raw	CNN	conv(1)	FC(3)	Relu	softmax	0.91 for S1; 0.87 for S2; 0.84 for S3; 0.69 for across subjects	DCGAN : 0.97 for S1, 0.93 for S2, 0.87 for S3; VAE: 0.73 across subjects	Video-Stimuli Dataset for augmentation, NAO Dataset for testing
[85]	SW	S	13	311.4h	6	2	57-63 &117-123	CF	STFT+GAN	GAN+CNN	conv(3) +conv(3)	FC(256)	sigmoid	softmax	0.60	0.72	Freiburg Hospital intracranial EEG dataset
[57]	SW	S	13	209h	22	2	47-53 & 97-103	CF	STFT+GAN	GAN+CNN	conv(3) +conv(3)	FC(256)	NA	NA	0.74	0.75	CHB-MIT database
[82]	SW	S	18	835 hours 1389 seizures	100	2	0.5-12.8	CF	spatio-temporal +information theory	CNN	conv(6)	Softmax	sigmoid	softmax	NA	0.97	Neonatal intensive Care Unit of Cork University(NICU) Maternity Hospital

[49]	SW	VT	7	varied	8	5	4.0-40	S	Raw	CNN	conv(2)	FC(5)	ELU	softmax	subject by subject	Increased 0.99	Due to ethical restrictions imposed by Korea University Institutional Review Board, data cannot be made publicly available
[83]	SW	MT	NA	288 trials/subject	NA	4	4-end	CF	spatial filter	CNN	conv(4)	FC(4)	Relu	softmax	NA	0.84	BCI competition IV-2a
[83]	SW	MI	NA	880 trials/subject	NA		4-end	S	Raw	NA	NA	NA	Relu	softmax	NA	0.95	High Gamma Dataset
[84]	SW	S	10	4090 trials	100	2 & 3	NA	S	Raw	CNN	conv(3)	FC(2)	NA	NA	NA	0.96	Bonn University
[84]	SW	S	10	4090 trials	100	2 & 3	NA	S	Raw	CNN	conv(3)	FC(2)	NA	NA	NA	0.95	Bonn University
[85]	SW	S	13	311.4h	6	2	57-63 &117-123	CF	STFT	CNN	conv(3)	FC(2)	NA	FC(10)	NA	NA	Freiburg Hospital intracranial EEG dataset
[85]	SW	S	13	209h	22	2	47-53 & 97-103	CF	STFT	CNN	conv(3)	FC(2)	Relu	softmax	NA	(+ 7% incraesed)	the Boston Children's Hospital-MIT scalp EEG dataset,
[85]	SW	S	2	15.5h	16	2	47-53 & 97-103	CF	STFT	CNN	conv(3)	FC(2)	Relu	softmax	0.89%%	0.889	American Epilepsy Society Seizure Prediction Challenge dataset(kaggle)
[86]	SW	IT	9	NA	22	2	8-36HZ	CF	spatial filter+PSD+covariance	CNN	NA	FC(100)	Relu	softmax	NA	NA	BCI- 2a and 2b competition IV

[86]	SW	IT	9	NA	3	2	8-36HZ	CF	spatial filter	CNN	NA	FC(100)	Relu	softmax	NA	0.75	BCI-2b
[87]	SW	SS	NA	15188trials	4	2-6 class	NA	S	Raw	CNN	conv(9)	FC(100)	NA	NA	0.829	0.857	Sleep EDF bank(Physionet database)
[46]	SW	SS	NA	62 nights	20	5	0-30HZ	CF	spatial filter	CNN	conv(3)	softmax	Relu	softmax	NA	0.812	MASS
[88]	SW	S	29	1037 minutes	2 & 18	2	0.5-70HZ	S	Raw	CNN	conv(4)	FC(2)	Relu	Relu	NA	0.93 for 2 channels 0.95 for 18 channels	KK women and children hospital
[89]	SW	MI	20	750 trials	3	2	2-60HZ	I	Spectrum	CNN	conv(3)	FC(2)	Relu	softmax	NA	0.84	Public dataset and BCI competitionIV dataset 2b
[90]	SW	S	24	980h	18	2	1-110HZ	CF	spatio-temporal +Graph theory+correlation	lstm	LSTM(1) & LSTM(2)	FC(2)	NA	NA	0.70	0.78	CHB-MIT database
[91]	SW	S	2	15min	varied	2	NA	S	Raw	CNN	conv(3)	FC(2)	NA	NA	NA	NA	kaggle
[92]	S	VT	12	NA	64	2	0-51HZ	S	Raw	conv	conv(3)	FC(2)	NA	NA	0.8399	0.8696	Hebrew university
[93]	S	MT	NA	3000trials	2	4	0.1-36HZ	CF	spatial filter	MLP	hid(3)	FC(40)	Relu	softmax	NA	0.82	MAHNOB HCI-Tagging database
[93]	S	MT	NA	3000trials	2	4	0.1-36HZ	CF	spatial filter	CNN	conv(2)	FC(10)	Relu	softmax	NA	0.93(6class)	MAHNOB HCI-Tagging database
[94]	S	SS	62	58600trials(30s)	20	5	0.3-100HZ	S	Raw	CNN	conv(4)	softmax	NA	NA	Arousal 0.69%, Valence 0.53%	Arousal 0.78%, Valence 0.73%	Advanced research in sleep medicine of the hopital du

																sacrecoeur de montreal
[94]	S	SS	20	41950trials(30s)	2	8	0.3-100HZ	S	Raw	CNN	conv(4)	softmax	NA	NA	0.776	0.793	Sleep-EDF
[95]	S	SS	62	494 hours	20	5	0-30HZ	CF	STFT	mixedNN	MLP+RNN	softmax	Relu	softmax	0.893	0.901	NA
[45]	S	SS	NA	5793trials	2	5	NA	S	Raw	CNN	conv(12)	FC(5)	sigmoid	softmax	0.60	0.74	SHHS(SHHS-1)
[96]	S	SS	121	NA	14	2	0.3-100HZ	I	spectrogram	CNN	conv(2)	FC(2)	NA	sigmoid	NA	0.70	Advanced research in sleep medicine of the hopital du sacrecoeur de montreal
[96]	S	SS	122	NA	14	2	0.3-100HZ	I	spectrogram	NN	FC(1)	FC(2)	Relu	softmax	NA	NA	Advanced research in sleep medicine of the hopital du sacrecoeur de montreal
[96]	S	SS	123	NA	14	2	0.3-100HZ	I	spectrogram	RNN	LSTM(3)	FC(2)	Relu	softmax	NA	NA	Advanced research in sleep medicine of the hopital du sacrecoeur de montreal
[97]	S	SS	20	43836 trials	2	5	0.3-35	CF	Energy, Power, Window deep belief window	LSTM	LSTM(2)	FC(5)	NA	softmax	0.846	0.855	Sleep-EDF and Sleep Apnea
[98]	S	S	23	NA	23	NA	0-49HZ	CF	spatio-temporal	CNN+LSTM	conv(4)+LSTM(1)	FC(64)	Relu	softmax	NA	NA	CHB_MIT

[99]	S	S	NA	10240 trials	NA	2	0.5-150Hz	I	Spectrogram	CNN	conv(4)	FC(2)	Relu	softmax	NA	0.99	www.upf.edu
[106]	FT	SS	NA	8h	16	6	0-13 HZ	S	Raw	CNN	conv(5)	softmax	sigmoid	softmax	0.72	0.75	CAPSLPDB
[42]	FT	MI	5	240Trials/subject	14	2	8HZ-30HZ	CF	spatio-temporal	CNN	conv(2)	FC(2)	Relu	softmax	NA	.9	Not available
[42]	FT	MI	5	240Trials/subject	14	2	8HZ-30HZ	CF	spatio-temporal	WNN	FC(2)	Relu	softmax	FC(2)	NA	0.85	Not available
[42]	FT	MI	1	280 trial	3	2	8HZ-30HZ	CF	spatio-temporal	CNN	conv(2)	FC(2)	Relu	softmax	0.88	0.82	BCI competition II, dataset III
[42]	FT	MI	1	280 trial	3	2	8HZ-30HZ	CF	spatio-temporal	WNN	FC(2)	Relu	softmax	FC(2)	0.88	0.84	BCI competition II, dataset III
[112]	O	ER	22	120 min	32	NA	4-100 HZ	CF	ICA/PCA	NN	hid(2)	Softmax	NA	NA	0.408	0.454	DEAP
[48]	O	MT	5	20 to 60/subject	1	2	1-30HZ	S	Raw	NN	hid(2)	NA	Relu	softmax	NA	0.97	NA
[48]	O	MT	5	20 to 60/subject	1	2	1-30HZ	S	Raw	NN	hid(2)	NA	sigmoid	sigmoid	NA	0.99 static 0.94 ambulatory conditions	NA
[48]	O	MT	5	20 to 60/subject	1	2	1-30HZ	S	Raw	NN	hid(2)	NA	ELU	softmax	NA	0.92	NA
[48]	O	MT	5	20 to 60/subject	1	2	1-30HZ	S	Raw	NN	hid(2)	NA	NA	NA	NA	NA	NA
[113]	O	S	155	~24 hours /subj	19	5	0-60 HZ	CF	Entropy	Hybrid(CNN+AE)	conv(6)	FC(5)	Relu	softmax	NA	0.814	Neuroscience ICU at Massachusetts
[114]	O	MI	1	280	3	2	NA	CF	STFT	CNN	conv(6)	FC(2)	Relu	softmax	NA	0.89	BCI competition II ,datset III
[114]	O	MI	9	400	3	2	NA	CF	STFT	CNN	conv(6)	FC(2)	Relu	softmax	NA	NA	BCI competition IV, dataset 2b
[115]	O	MI	9	4session 72trials/session	3	4	7-30Hz	S	Raw	CNN+LSTM	varied	varied	varied	varied	NA	Overal 5.3%	BCI competition IV, dataset 2a

															improved accuracy	
[116]	O	ER	32	40minutes/subject	32	2	4-45Hz	CF	Wavelet	CNN	conv(2)	FC(2)	Relu	Relu	Arousal: 0.86 Valence: 0.88	Arousal: 0.91 Valence: 0.91	DEAP
[109]	RS	ER	32	23040trials	32	4	NA	S	Wavelet	SAE	Unspecified	Unspecified	Exponential Linear Units	softmax	NA	0.6875	DEAP
[89]	RS	S	23	Varied for each subject	18	2	5-50Hz	CF	Wavelet	CNN	conv(2)	FC(2)	Leaky Relu	softmax	NA	0.9	CHB-MIT
[111]	RS	MI	9	6520 trials	3	2	4-32Hz	S	Raw	CNN	conv(2)	FC(2)	ELU	softmax	0.86	0.87	BCI competition IV, dataset 2b

References

[1] Yannick, R., et al., *Deep learning-based electroencephalography analysis: a systematic review*. arXiv preprint arXiv:1901.05498, 2019.

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

[3] Bigdely-Shamlo, N., et al., *The PREP pipeline: standardized preprocessing for large-scale EEG analysis.* Frontiers in neuroinformatics, 2015. **9**: p. 16.

[4] Jas, M., et al., *Autoreject: automated artifact rejection for MEG and EEG data.* NeuroImage, 2017. **159**: p. 417-429.

[5] Cole, S.R. and B. Voytek, *Cycle-by-cycle analysis of neural oscillations.* bioRxiv, 2018: p. 302000.

[6] Gramfort, A., et al., *Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations.* NeuroImage, 2013. **70**: p. 410-422.

[7] Shanechi, M.M., *Brain–machine interfaces from motor to mood.* Nature neuroscience, 2019. **22**(10): p. 1554-1564.

[8] Sani, O., et al., *Neural Decoding and Control of Mood to Treat Neuropsychiatric Disorders.* Biological Psychiatry, 2020. **87**(9): p. S95-S96.

[9] Hively, L., V. Protopopescu, and P. Gailey, *Timely detection of dynamical change in scalp EEG signals.* Chaos: An Interdisciplinary Journal of Nonlinear Science, 2000. **10**(4): p. 864-875.

[10] Subasi, A. and M.I. Gursoy, *EEG signal classification using PCA, ICA, LDA and support vector machines.* Expert systems with applications, 2010. **37**(12): p. 8659-8666.

[11] Sanz-García, A., et al., *Potential EEG biomarkers of sedation doses in intensive care patients unveiled by using a machine learning approach.* Journal of neural engineering, 2019. **16**(2): p. 026031.

[12] He, H. and E.A. Garcia, *Learning from imbalanced data.* IEEE Transactions on knowledge and data engineering, 2009. **21**(9): p. 1263-1284.

[13] Lotte, F., et al., *A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update.* Journal of neural engineering, 2018. **15**(3): p. 031005.

[14] Chollet, F., *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. 2018: MITP-Verlags GmbH & Co. KG.

[15] Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. 2016: MIT press.

[16] Farabet, C., et al., *Learning hierarchical features for scene labeling.* IEEE transactions on pattern analysis and machine intelligence, 2013. **35**(8): p. 1915-1929.

[17] Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.

[18] Tompson, J.J., et al. *Joint training of a convolutional network and a graphical model for human pose estimation*. in *Advances in neural information processing systems*. 2014.

[19] Hinton, G., et al., *Deep neural networks for acoustic modeling in speech recognition.* IEEE Signal processing magazine, 2012. **29**.

[20] Mikolov, T., et al. *Strategies for training large scale neural network language models*. in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. 2011. IEEE.

[21] Sainath, T.N., et al. *Deep convolutional neural networks for LVCSR*. in *2013 IEEE international conference on acoustics, speech and signal processing*. 2013. IEEE.

[22] LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436.

[23] Craik, A., Y. He, and J.L. Contreras-Vidal, *Deep learning for electroencephalogram (EEG) classification tasks: a review.* Journal of neural engineering, 2019. **16**(3): p. 031001.

[24] Page, A., C. Shea, and T. Mohsenin. *Wearable seizure detection using convolutional neural networks with transfer learning*. in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2016. IEEE.

[25] Li, J., et al., *Multisource transfer learning for cross-subject eeg emotion recognition.* IEEE transactions on cybernetics, 2019.

[26] Fahimi, F., et al., *Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI.* Journal of neural engineering, 2019. **16**(2): p. 026007.

[27] Perez, L. and J. Wang, *The effectiveness of data augmentation in image classification using deep learning.* arXiv preprint arXiv:1712.04621, 2017.

[28] Wang, F., et al. *Data augmentation for eeg-based emotion recognition with deep convolutional neural networks*. in *International Conference on Multimedia Modeling*. 2018. Springer.

[29] Zhang, C., et al., *Understanding deep learning requires rethinking generalization.* arXiv preprint arXiv:1611.03530, 2016.

[30] Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting.* The journal of machine learning research, 2014. **15**(1): p. 1929-1958.

[31] Srivastava, N., *Improving neural networks with dropout.* University of Toronto, 2013. **182**(566): p. 7.

[32] Shorten, C. and T.M. Khoshgoftaar, *A survey on image data augmentation for deep learning.* Journal of Big Data, 2019. **6**(1): p. 60.

[33] Kukačka, J., V. Golkov, and D. Cremers, *Regularization for deep learning: A taxonomy.* arXiv preprint arXiv:1710.10686, 2017.

[34] Deepa, V., *Investigating the performance improvement by sampling techniques in EEG data.* International Journal on Computer Science and Engineering, 2010. **2**(6).

[35] Bhattacharyya, A. and R.B. Pachori, *A multivariate approach for patient-specific EEG seizure detection using empirical wavelet transform.* IEEE Transactions on Biomedical Engineering, 2017. **64**(9): p. 2003-2015.

[36] Romaissa, D., M. El Habib, and M.A. Chikh. *Epileptic Seizure Detection from Imbalanced EEG signal*. in *2019 International Conference on Advanced Electrical Engineering (ICAEE)*. 2019. IEEE.

[37] LeCun, Y., et al., *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 1998. **86**(11): p. 2278-2324.

[38] Barry, R.J. and H.C. Beh, *EEG correlates of the afterimage of visual stimulation.* Psychophysiology, 1976. **13**(1): p. 75-80.

[39] Del Re, E., *Bandpass signal filtering and reconstruction through minimum-sampling-rate digital processing.* Alta Frequenza, 1978. **47**(9): p. 675-678.

[40] Truong, N.D., et al., *Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram.* Neural Networks, 2018. **105**: p. 104-111.

[41] Dümpelmann, M., *Early seizure detection for closed loop direct neurostimulation devices in epilepsy.* Journal of neural engineering, 2019. **16**(4): p. 041001.

[42] Zhang, Z., et al., *A novel deep learning approach with data augmentation to classify motor imagery signals.* IEEE Access, 2019. **7**: p. 15945-15954.

[43] Malhotra, R.K. and A.Y. Avidan, *Sleep stages and scoring technique.* Atlas of sleep medicine, 2013: p. 77-99.

[44] Moser, D., et al., *Sleep classification according to AASM and Rechtschaffen & Kales: effects on sleep scoring parameters.* Sleep, 2009. **32**(2): p. 139-149.

[45] Sors, A., et al., *A convolutional neural network for sleep stage scoring from raw single-channel EEG.* Biomedical Signal Processing and Control, 2018. **42**: p. 107-114.

[46] Chambon, S., et al., *A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2018. **26**(4): p. 758-769.

[47] Kuanar, S., et al. *Cognitive Analysis of Working Memory Load from EEG, by a Deep Recurrent Neural Network*. in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018. IEEE.

[48] Sakai, A., Y. Minoda, and K. Morikawa. *Data augmentation methods for machine-learning-based classification of bio-signals*. in *2017 10th Biomedical Engineering International Conference (BMEiCON)*. IEEE.

[49] Kwak, N.-S., K.-R. Müller, and S.-W. Lee, *A convolutional neural network for steady state visual evoked potential classification under ambulatory environment.* PloS one, 2017. **12**(2): p. e0172578.

[50] Bashivan, P., et al., *Learning representations from EEG with deep recurrent-convolutional neural networks.* arXiv preprint arXiv:1511.06448, 2015.

[51] Yin, Z. and J. Zhang, *Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights.* Neurocomputing, 2017. **260**: p. 349-366.

[52] Yin, Z. and J. Zhang, *Cross-session classification of mental workload levels using EEG and an adaptive deep learning model.* Biomedical Signal Processing and Control, 2017. **33**: p. 30-47.

[53] Hussein, R., et al., *Epileptic seizure detection: A deep learning approach.* arXiv preprint arXiv:1803.09848, 2018.

[54] Salama, E.S., et al., *EEG-based emotion recognition using 3D convolutional neural networks.* Int. J. Adv. Comput. Sci. Appl, 2018. **9**(8): p. 329-337.

[55] Parvan, M., et al. *Transfer Learning based Motor Imagery Classification using Convolutional Neural Networks*. in *2019 27th Iranian Conference on Electrical Engineering (ICEE)*. 2019. IEEE.

[56] Li, Y., et al., *A Channel-Projection Mixed-Scale Convolutional Neural Network for Motor Imagery EEG Decoding.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2019.

[57] Goodfellow, I., et al. *Generative adversarial nets*. in *Advances in neural information processing systems*. 2014.

[58] Zhang, H., et al. *Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[59] Bousmalis, K., et al. *Unsupervised pixel-level domain adaptation with generative adversarial networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[60] Antoniou, A., A. Storkey, and H. Edwards, *Data augmentation generative adversarial networks.* arXiv preprint arXiv:1711.04340, 2017.

[61] Zhang, Q. and Y. Liu, *Improving brain computer interface performance by data augmentation with conditional Deep Convolutional Generative Adversarial Networks.* arXiv preprint arXiv:1806.07108, 2018.

[62] Piplani, T., N. Merill, and J. Chuang, *Faking it, Making it: Fooling and Improving Brain-Based Authentication with Generative Adversarial Networks.*

[63] Zhang, X., et al., *DADA: Deep adversarial data augmentation for extremely low data regime classification.* arXiv preprint arXiv:1809.00981, 2018.

[64] Schlögl, A., *Outcome of the BCI-competition 2003 on the Graz data set.* Berlin, Germany: Graz University of Technology, 2003.

[65] Hartmann, K.G., R.T. Schirrmeister, and T. Ball, *EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals.* arXiv preprint arXiv:1806.01875, 2018.

[66] Salimans, T., et al. *Improved techniques for training gans*. in *Advances in neural information processing systems*. 2016.

[67] Heusel, M., et al. *Gans trained by a two time-scale update rule converge to a local nash equilibrium*. in *Advances in Neural Information Processing Systems*. 2017.

[68] Rabin, J., et al. *Wasserstein barycenter and its application to texture mixing*. in *International Conference on Scale Space and Variational Methods in Computer Vision*. 2011. Springer.

[69] Luo, Y. and B.-L. Lu. *EEG data augmentation for emotion recognition using a conditional Wasserstein GAN*. in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018. IEEE.

[70] Zheng, W.-L. and B.-L. Lu, *Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks.* IEEE Transactions on Autonomous Mental Development, 2015. **7**(3): p. 162-175.

[71] Koelstra, S., et al., *Deap: A database for emotion analysis; using physiological signals.* IEEE transactions on affective computing, 2012. **3**(1): p. 18-31.

[72] Luo, Y., L.-Z. Zhu, and B.-L. Lu. *A GAN-Based Data Augmentation Method for Multimodal Emotion Recognition*. in *International Symposium on Neural Networks*. 2019. Springer.

[73] Wei, Z., et al., *Automatic epileptic EEG detection using convolutional neural network with improvements in time-domain.* Biomedical Signal Processing and Control, 2019. **53**: p. 101551.

[74] Chang, S. and H. Jun, *Hybrid deep-learning model to recognise emotional responses of users towards architectural design alternatives.* Journal of Asian Architecture and Building Engineering, 2019. **18**(5): p. 381-391.

[75] Yang, B., et al. *A Framework on Optimization Strategy for EEG Motor Imagery Recognition*. in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019. IEEE.

[76] Panwar, S., et al. *Generating EEG signals of an RSVP Experiment by a Class Conditioned Wasserstein Generative Adversarial Network*. in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 2019. IEEE.

[77] Gulrajani, I., et al. *Improved training of wasserstein gans*. in *Advances in neural information processing systems*. 2017.

[78] Touryan, J., et al., *Estimating endogenous changes in task performance from EEG.* Frontiers in neuroscience, 2014. **8**: p. 155.

[79] Radford, A., L. Metz, and S. Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks.* arXiv preprint arXiv:1511.06434, 2015.

[80] Kingma, D.P. and M. Welling, *Auto-encoding variational bayes.* arXiv preprint arXiv:1312.6114, 2013.

[81] Aznan, N.K.N., et al. *Using variable natural environment brain-computer interface stimuli for real-time humanoid robot navigation*. in *2019 International Conference on Robotics and Automation (ICRA)*. 2019. IEEE.

[82] O'Shea, A., et al. *Neonatal seizure detection using convolutional neural networks*. in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. 2017. IEEE.

[83] Schirrmeister, R.T., et al., *Deep learning with convolutional neural networks for EEG decoding and visualization.* Human brain mapping, 2017. **38**(11): p. 5391-5420.

[84] Ullah, I., M. Hussain, and H. Aboalsamh, *An automated system for epilepsy detection using EEG brain signals based on deep learning approach.* Expert Systems with Applications, 2018. **107**: p. 61-71.

[85] Truong, N.D., et al., *Semi-supervised Seizure Prediction with Generative Adversarial Networks.* arXiv preprint arXiv:1806.08235, 2018.

[86] Majidov, I. and T. Whangbo, *Efficient Classification of Motor Imagery Electroencephalography Signals Using Deep Learning Methods.* Sensors, 2019. **19**(7): p. 1736.

[87] Mousavi, Z., et al., *Deep convolutional neural network for classification of sleep stages from single-channel EEG signals.* Journal of neuroscience methods, 2019: p. 108312.

[88] Avcu, M.T., Z. Zhang, and D.W.S. Chan. *Seizure Detection Using Least Eeg Channels by Deep Convolutional Neural Network*. in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019. IEEE.

[89] Tayeb, Z., et al., *Validating deep neural networks for online decoding of motor imagery movements from EEG signals.* Sensors, 2019. **19**(1): p. 210.

[90] Tsiouris, K.M., et al., *A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals.* Computers in biology and medicine, 2018. **99**: p. 24-37.

[91] Tang, Y., S. Wada, and K. Yoshihara. *Failure Prediction with Adaptive Multi-scale Sampling and Activation Pattern Regularization*. in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2017. IEEE.

[92] Manor, R. and A.B. Geva, *Convolutional neural network for multi-category rapid serial visual presentation BCI.* Frontiers in computational neuroscience, 2015. **9**: p. 146.

[93] Drouin-Picaro, A. and T.H. Falk. *Using deep neural networks for natural saccade classification from electroencephalograms*. in *2016 IEEE EMBS International Student Conference (ISC)*. 2016. IEEE.

[94] Supratak, A., et al., *DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2017. **25**(11): p. 1998-2008.

[95] Dong, H., et al., *Mixed neural network approach for temporal sleep stage classification.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2017. **26**(2): p. 324-333.

[96] Ruffini, G., et al., *Deep learning using EEG spectrograms for prognosis in idiopathic rapid eye movement behavior disorder (RBD).* bioRxiv, 2018: p. 240267.

[97] Sun, C., et al., *A Two-Stage Neural Network for Sleep Stage Classification Based on Feature Learning, Sequence Learning, and Data Augmentation.* IEEE Access, 2019. **7**: p. 109386-109397.

[98] Thodoroff, P., J. Pineau, and A. Lim. *Learning robust features using deep learning for automatic seizure detection*. in *Machine learning for healthcare conference*. 2016.

[99] Sengur, A., et al., *Neutrosophic similarity score-based entropy measure for focal and nonfocal electroencephalogram signal classification*, in *Neutrosophic Set in Medical Image Analysis*. 2019, Elsevier. p. 247-268.

[100] Kubat, M. and S. Matwin. *Addressing the curse of imbalanced training sets: one-sided selection*. in *Icml*. 1997. Citeseer.

[101] Japkowicz, N. *The class imbalance problem: Significance and strategies*. in *Proc. of the Int'l Conf. on Artificial Intelligence*. 2000. Citeseer.

[102] Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.

[103] Vega, R., *Prediction of seizures using canine EEG data and deep learning techniques: a comparison between convolutional, and recurrent neural networks.*

[104] Wang, S., et al. *Training deep neural networks on imbalanced data sets*. in *2016 international joint conference on neural networks (IJCNN)*. 2016. IEEE.

[105] Moniz, N., P. Branco, and L. Torgo. *Resampling strategies for imbalanced time series*. in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2016. IEEE.

[106] Schwabedal, J.T., et al., *Addressing Class Imbalance in Classification Problems of Noisy Signals by using Fourier Transform Surrogates.* arXiv preprint arXiv:1806.08675, 2018.

[107] Ruffini, G., et al., *Deep learning with EEG spectrograms in rapid eye movement behavior disorder.* bioRxiv, 2018: p. 240267.

[108] Shamwell, J., et al. *Single-trial EEG RSVP classification using convolutional neural networks*. in *Micro-and Nanotechnology Sensors, Systems, and Applications VIII*. 2016. International Society for Optics and Photonics.

[109] Said, A.B., et al. *Multimodal deep learning approach for joint EEG-EMG data compression and classification*. in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*. 2017. IEEE.

[110] Zhang, Y., et al., *Epilepsy seizure prediction on eeg using common spatial pattern and convolutional neural network.* IEEE Journal of Biomedical and Health Informatics, 2019.

[111] Dai, G., et al., *HS-CNN: a CNN with hybrid convolution scale for EEG motor imagery classification.* Journal of neural engineering, 2020. **17**(1): p. 016025.

[112] Frydenlund, A. and F. Rudzicz. *Emotional affect estimation using video and EEG data in deep neural networks*. in *Canadian Conference on Artificial Intelligence*. 2015. Springer.

[113] Deiss, O., et al., *HAMLET: Interpretable Human And Machine co-LEarning Technique.* arXiv preprint arXiv:1803.09702, 2018.

[114] Shovon, S.T.H., et al. *Classification of Motor Imagery EEG Signals with multi-input Convolutional Neural Network by augmenting STFT*. in *5th International Conference on Advances in Electrical Engineering (ICAEE). IEEE*. 2019.

[115] Freer, D. and G.-Z. Yang, *Data augmentation for self-paced motor imagery classification with C-LSTM*. Journal of neural engineering, 2019.

[116] Mokatren, L.S., et al., *Improved EEG Classification by factoring in sensor topography*. arXiv preprint arXiv:1905.09472, 2019.

[117] Raza, H., et al., *Adaptive learning with covariate shift-detection for motor imagery-based brain–computer interface.* Soft Computing, 2016. **20**(8): p. 3085-3096.

[118] Gaur, P., et al. *An empirical mode decomposition based filtering method for classification of motor-imagery EEG signals for enhancing brain-computer interface*. in *2015 International Joint Conference on Neural Networks (IJCNN)*. 2015. IEEE.

[119] Freer, D. and G.-Z. Yang, *Data augmentation for self-paced motor imagery classification with C-LSTM*. Journal of Neural Engineering, 2020. **17**(1): p. 016041.