Virtual EEG-electrodes: Convolutional neural networks as a method for upsampling or restoring channels

Mats Svantesson^{a,b,c,f,*}, Håkan Olausson^{a,b,f}, Anders Eklund^{c,d,e}, Magnus Thordstein^{a,c,f}

^a Department of Clinical Neurophysiology, University Hospital of Linköping, Sweden

^b Center for Social and Affective Neuroscience, Linköping University, Sweden

^c Center for Medical Image Science and Visualization, Linköping University, Sweden

^d Department of Biomedical Engineering, Linköping University, Sweden

^e Division of Statistics & Machine Learning, Department of Computer and Information Science, Linköping University, Sweden

^f Department of Biomedical and Clinical Sciences, Linköping University, Sweden

ARTICLE INFO

Keywords:

Deep learning
Convolutional neural networks
Electroencephalography
Signal reconstruction
Spatial upsampling
Spherical spline interpolation

ABSTRACT

Background: In clinical practice, EEGs are assessed visually. For practical reasons, recordings often need to be performed with a reduced number of electrodes and artifacts make assessment difficult. To circumvent these obstacles, different interpolation techniques can be utilized. These techniques usually perform better for higher electrode densities and values interpolated at areas far from electrodes can be unreliable. Using a method that learns the statistical distribution of the cortical electrical fields and predicts values may yield better results.

New Method: Generative networks based on convolutional layers were trained to upsample from 4 or 14 channels or to dynamically restore single missing channels to recreate 21-channel EEGs. 5,144 h of data from 1,385 subjects of the Temple University Hospital EEG database were used for training and evaluating the networks.

Comparison with Existing Method: The results were compared to spherical spline interpolation. Several statistical measures were used as well as a visual evaluation by board certified clinical neurophysiologists. Overall, the generative networks performed significantly better. There was no difference between real and network generated data in the number of examples assessed as artificial by experienced EEG interpreters whereas for data generated by interpolation, the number was significantly higher. In addition, network performance improved with increasing number of included subjects, with the greatest effect seen in the range 5–100 subjects.

Conclusions: Using neural networks to restore or upsample EEG signals is a viable alternative to spherical spline interpolation.

1. Introduction

In clinical routine, EEG is assessed by visual inspection by a trained human interpreter. Ongoing work strives to develop methods for automating the analysis of EEG (Acharya et al., 2013; Roy et al., 2019). However, it is likely that the analysis will continue to be visual for the foreseeable future and that there will be a long period of transition, where visual and automated analysis will be performed in parallel, before the analysis becomes fully automated. Developing methods to enhance the visual assessment is thus motivated.

EEG is used increasingly for long term monitoring of cerebral function, e.g., to detect seizures or status epilepticus (Kubota et al., 2018). The number of scalp electrodes used in this setting is often reduced compared to a standard EEG (e.g., 4 instead of 21 electrodes), due to

limitations in resources and practical problems with maintaining signal quality for longer time periods (Hera et al., 2017). A low electrode density means a low spatial resolution, making the recordings harder to assess compared to standard EEGs since identification and classification of wave phenomena often depend on spatial patterns (Schomer et al., 2018). Comparing the sensitivity of reduced and full EEG montages has its difficulties. Generalizing results from one reduced montage to another may not hold since the sensitivity for different EEG patterns may vary depending on the electrode positions (Bennis et al., 2017; Foldvary et al., 2000; Kolls and Husain, 2007; Westover et al., 2020). Interrater variability and asymmetric comparisons between reduced and full montages with regard to the amount of EEG data and ancillary information (e.g., patient history, medication, radiological findings) contribute to the problem of establishing 'gold standards' used to assess

* Corresponding author at: Department of Clinical Neurophysiology, University Hospital of Linköping, 58185 Linköping, Sweden.

E-mail address: mats.svantesson@liu.se (M. Svantesson).

<https://doi.org/10.1016/j.jneumeth.2021.109126>

Received 16 July 2020; Received in revised form 21 January 2021; Accepted 4 March 2021

Available online 9 March 2021

0165-0270/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

sensitivity (Westover et al., 2020). Studies thus differ in methodology and have varying results, and even though there are examples of high sensitivities for certain EEG patterns in some studies (Backman et al., 2020; Gururangan et al., 2018; Ma et al., 2018; Pati et al., 2017; Vanherpe and Schrooten, 2017), the diagnostic value and limitations of reduced montages have not conclusively been determined.

Common methods used for scalp potential interpolation are nearest-neighbor and splines (Fletcher et al., 1996). For instance, the commercial EEG analysis software Curry (Compumedics Neuroscan, Dresden, Germany) has a nearest-neighbor implementation, the Python library MNE (Gramfort et al., 2013) and Matlab toolbox EEGLAB (Delorme and Makeig, 2004) use spherical splines as their default methods, and the FieldTrip toolbox (Oostenveld et al., 2011) uses the weighted average of nearest-neighbors as its default with the option to use spherical splines. Spline techniques seem to perform better than nearest-neighbor with a tendency for spherical splines to perform best (Perrin et al., 1987, 1989; Soong et al., 1993). Bilinear and bicubic interpolation are other common methods used for interpolation (Koles and Paranjape, 1988). We found few studies systematically comparing these with splines methods, but thin plate splines perform better than bilinear interpolation in topographic mapping (Satherley et al., 1996). It has been concluded that ‘...adequate electrode density is more important than the method used’ and that there is a risk of large interpolation errors in areas distant to electrodes (Fletcher et al., 1996). It is also claimed that the electrode density used clinically, i.e., the international 10–20 system (Jasper, 1958), is too low for interpolation (Soong et al., 1993).

When deep learning applications are used for EEG analysis, convolutional neural networks (CNNs) are most common, used in 40 % of the papers (Roy et al., 2019). Some of these studies investigate the generation of synthetic EEG data, with the majority of networks implemented as generative adversarial networks (GANs). Most of the studies concern the generation of synthetic signals for data augmentation and produce realistic signals and improve classification of data (Pascual et al., 2019; Aznan et al., 2019; Zhang and Liu, 2018; Hartmann et al., 2018; Luo and Lu, 2018; Qiu and Zhao, 2018).

Few previous studies relate to upsampling. In one such study, Luo et al. (2020) showed an improvement in classification tasks using Wasserstein GANs (WGANs) for temporal upsampling. In another, Corley and Huang (2019) demonstrated a reduction of the mean absolute error by one to two orders of magnitude compared to cubic interpolation, using WGANs to upsample from 8 or 16–32 channels. They used the training dataset ‘V’ of the Berlin Brain Computer Interface Competition III (Millan, 2004) consisting of 2,141 seconds of 32-channel EEG data from 3 subjects. In a third study, Kwon et al. (2019) used CNNs to recreate 64-channel EEG from as little as 4 channels. Compared to linear interpolation, the mean square error they found was lower and the correlation with the original data was stronger. In addition, they found that source localization was fairly adequate using recreated signals whereas it failed with interpolated data. They used simulated EEG data (640 examples) and experimental EEG data from an auditory task (596 examples from 5 subjects) for training.

It is usually necessary to train deep neural networks with large amounts of data (Sun et al., 2017). For EEG data, it may also be important to train using an adequate number of subjects, for results to be generalizable. This has not been thoroughly investigated, but Völker et al. (2018) demonstrated an improvement in classification accuracy with increasing number of subjects, most notably when more than 15 were used, in experiments using 1,500 ms examples of 128-channel EEG data from up to 30 subjects in a cognitive task (the number of examples was not reported).

In this study we further investigated the use of CNNs to upsample the electrode density or to restore signals of EEG and compare the results to a standard method used in common software packages for EEG processing, spherical splines. We evaluated the signal quality visually as well as with several objective measures to assess the closeness to real signals. As discussed above, previous studies (Corley and Huang, 2019;

Kwon et al., 2019) have used a limited amount of data and few subjects, but here we used a large data set of real EEG and also tested to what extent training the networks with different number of subjects affected the results.

2. Materials and methods

2.1. Software and hardware

The computers used the operating system Ubuntu version 18.04.2 LTS. The networks were developed in Python (version 3.6.5) using the API Keras (version 2.2.4) and programming module TensorFlow (version 1.10.1). Random seed was set to 12,345 for the libraries ‘random’ and ‘numpy.random’.

The Python library ‘pyEDFlib’ (Nahrstaedt and Lee-Messer, 2019) was used to extract EEG data. Extracting subject information, i.e., age and gender, was not possible with the ‘pyEDFlib’, so MATLAB R2018a was used with a downloaded script (Shapkin, 2012). The MNE-Python library (Gramfort et al., 2013) was used to perform spherical spline interpolation.

Two computers were used for the study, equipped with: 1) Intel Xeon E5 1620V4 Quad Core 3.5 GHz processor, 96 GB 2,400 MHz RAM, and Nvidia Quadro P5000 GPU; 2) Intel Core i7-6800 K Hexa Core 3.4 GHz, 96 GB 2,133 MHz RAM, Nvidia Quadro P6000, and Nvidia Titan X.

The code is available at: <https://github.com/Svanteberg/Virtual-EEG-electrodes>.

2.2. EEG data

The EEG data used for this study was acquired from the published database created at the Temple University Hospital (TUH), Philadelphia (Obeid and Picone, 2016). The TUH EEG Corpus (v1.1.0) with average reference was used (downloaded during 17–21 January 2019).

The recording electrodes are positioned according to the international 10–20 system (Jasper, 1958). A full EEG montage here was defined as consisting of 21 electrodes Fp1, F7, T3, T5, Fp2, F8, T4, T6, F3, C3, P3, O1, F4, C4, P4, O2, A1, A2, Fz, Cz, and Pz; here given in the order used for visual analysis. Many recordings include more electrodes but only those listed here were used. The data was not re-referenced. All recordings are in European Data Format (EDF).

The database contains subjects of all ages, but there are relatively few non-adults. Since the cortical activity of children can differ (Pearl et al., 2018), only subjects of 18 years or older were included. To increase the possibility that each used recording contained appreciable variations of the signals, the duration of each had to be at least 300 s to be included. This limit was arbitrary. The recordings of the database are sampled at different rates. To have matching sample rates, only recordings sampled at 256 Hz were used. Based on these criteria a total of 1,385 (female/male: 751/634) subjects with 11,163 (female/male: 5,648/5,515) recordings, corresponding to 5,144 h, was extracted from the database. Ages varied from 18 to 95 years with a mean age of 51 ± 18 years. The data contained normal and pathological as well as wake and sleep recordings of unknown distributions. All original data is unfiltered.

2.2.1. Data pre-processing

All data was high pass filtered at 0.3 Hz and low pass filtered at 40 Hz using second-degree Butterworth filters; the bandwidth was chosen according to our clinical preference. A 60 Hz notch filter was used to remove residual AC-noise. Filtering was applied with zero phase shift.

The data were divided into a training, validation and test set (80, 10 and 10 percent distribution, respectively). To ensure that the sets were disjoint with regard to subject, each subject was randomly assigned to one of the sets according to the distribution, resulting in 1,114, 129 and 142 subjects for each data set.

The training data consisted of 1,114 subjects providing, in total, 3,976 h of data. To evaluate the effect of the number of subjects used, networks were also trained with subsets of 5, 15, 30, 50, 100 and 500 randomly chosen subjects. This resulted in subsets consisting of 11, 37, 152, 159, 414 and 1,694 h of data, respectively.

Data examples with a duration of 10 s were generated randomly during training, validation and testing (see section 2.3.3); to avoid very large artifacts, drawn examples with maximum absolute amplitude over 500 μ V were discarded. The first and last 40 s of each recording was also discarded since these parts sometimes had more artifacts or did not contain any cortical signal (usually a low frequency square wave). This time interval was arbitrarily chosen from visual assessment of recordings, did not exclude all epochs without cortical signal and also resulted in discarding epochs with good data. No other measures were taken to reduce artifacts.

Each example was normalized to have a standard deviation of one. Since the data was high pass filtered, the mean was close to zero on average. It was also evaluated to use no normalization or to normalize each example to a maximal amplitude of one. The chosen option performed best, by giving better estimation of the amplitudes. The other approaches tended to more often give an underestimation of the amplitudes, where using no normalization was the least favorable.

2.2.2. Ethical considerations

The HIPAA Privacy Rule was followed when the TUH EEG Corpus was created (Obeid and Picone, 2016). Since the used data was from a published database with no subject information and therefore no way of tracing the data back to the subjects, no further approval from an ethics committee was deemed necessary for this study.

2.3. The networks

2.3.1. Architecture

In this study, three generative networks (GNs) were trained for processing 10 s examples of EEG data in slightly different ways. The first network (GN1) upsampled from 4 to 21 electrodes, the second network (GN2) upsampled from 14 to 21 electrodes, and the third network (GN3) had all 21 electrodes as inputs but with one of them randomly blocked and generated the blocked electrode as output. The electrodes used as inputs were not recreated meaning that GN1 had 4 inputs and 17 outputs, GN2 had 14 inputs and 7 outputs and GN3 had 21 inputs and 1 output (Table 1). The networks shared the same general architecture, the only essential difference was in the number of input and output channels.

GN1 represents a realistic scenario since four-electrode montages are used in some intensive care units or during anesthesia (Friberg et al., 2013; Pati et al., 2017; Tian et al., 2012). It is an underdetermined problem since the number of inputs are fewer than the outputs with some of the latter being spatially distant (Fig. 1). The specific input electrodes were chosen since they are the choice of preference at our clinic, as well as many other units. GN2 was trained as a comparison. In

this case there were more inputs than outputs, the electrodes had a spatially uniform distribution, and recreated electrodes had input electrodes as nearest-neighbors (Fig. 1) providing more favorable conditions for the recreation of the signals.

GN3 differed in its function since it in addition to having to learn to recreate signals, also had to learn to detect which electrode was missing. This type of network could be used to replace single missing or poorly performing channels, but the intent was also to test the ability of the networks to perform several tasks and the possibility for automatization. Ideally, one would like to have a network that can process multiple different aspects of an EEG, e.g., upsample, remove artifacts and detect events, and includes an option to do so automatically.

The networks were based on convolutional and deconvolutional layers, and except for the last layer, all layers were followed by a leaky rectifying linear unit (Maas et al., 2013). Convolutions and deconvolutions were performed separately for temporal and spatial dimensions.

The networks had three principal parts: First part, a temporal encoder consisting of four convolutional layers, using strides of two and doubling the number of filters for each layer. Second, a spatial analysis and upsampling performed by convolution with a filter size corresponding to the number of input electrodes, and then deconvoluting to the right number of output electrodes. Third, a temporal decoder of four deconvolutional layers with strides of two and a finishing convolutional layer to merge the filters resulting in the desired size (see Table 2). With a duration of 10 s, this corresponded to 2560 samples per electrode for in- and output. GN1 and GN2 had 5,962,049 and GN3 had 6,224,193 parameters, all trainable.

The rationale for this architecture was partly due to the data being spatially discontinuous since spatially invariant features in three dimensions may become disrupted when the data is arranged into a one-dimensional array rendering performing convolutions less effective. The intent was for each electrode to first be analyzed for (temporal) waveform features in the encoder section. Thereafter, the spatial analysis was performed over all electrodes rendering the analysis independent of the relative positions of the electrodes in the spatial dimension. After spatial upsampling, it was assumed that the decoder section reconstructed the signals by assembling waveform features.

2.3.2. Hyperparameters

In this work, we developed and tested a method for generating artificial EEG data intended to be used to aid visual analysis. The visual appearance was thus paramount. A problem when evaluating the quality of generated data is that there are no good objective measures that capture all aspects of the visual impression. For instance, the appearance of two different recreations of a signal having the same mean absolute error may be very different since the measure does not reflect the distribution of the errors. For visual analysis, the recreation of a correct waveform morphology may be more important than a correct amplitude. Furthermore, if a network generates low-amplitude or local artifacts, this may not substantially impact the measures but may be easily spotted visually (Fig. 2). This means that when choosing parameters, blindly settling for the network with the lowest training/validation loss may result in discarding a better option. A visual assessment of course is subjective, by nature has a qualitative rather than a quantitative character, and usually can only be performed on a limited amount of data. However, until reliable quantitative measures have been developed, we believe that the visual assessment is a relevant and necessary part. Hence, the comparison and selection of networks and parameters were mostly based on the overall validation loss, the visual appearance, and practical considerations (e.g., processing times).

Different combinations of number of layers (2–7), filter sizes (3, 5, 7, 11), and strides (1–2) were tested. For more than four layers, strides of one were not tested due to processing times. The validation losses were about the same for the nets (5.5–6.0 μ V), but with a small tendency for lower values with increasing number of layers. The visual appearance was assessed by a specialist in clinical neurophysiology (MS). The

Table 1
Input and output electrodes of the three networks.

Network	Input electrodes	Output electrodes
GN1	F3, P3, F4, P4	Fp1, F7, T3, T5, Fp2, F8, T4, T6, C3, O1, C4, O2, A1, A2, Fz, Cz, Pz
GN2	Fp1, F7, T5, Fp2, F8, T6, C3, O1, C4, O2, A1, A2, Fz, Pz	T3, T4, F3, P3, F4, P4, Cz
GN3	Fp1, F7, T3, T5, Fp2, F8, T4, T6, F3, C3, P3, O1, F4, C4, P4, O2, A1, A2, Fz, Cz, Pz*	One randomly chosen of the inputs

The given electrode order is the same as the order they are represented in the data. *) The randomly chosen electrode is blocked as input and replaced with low amplitude noise.

Fig. 1. Illustrations of electrode positions for in- and output signals for the GN1 (left) and GN2 (right) networks.

Table 2

Overview of the generative network.

	Layer type	Filters	Filter size	Strides	Padding	Activation
Temporal encoder	Conv2D	32	(1,3)	(1,2)	Same	LeakyReLU
	Conv2D	64	(1,3)	(1,2)	Same	LeakyReLU
	Conv2D	128	(1,3)	(1,2)	Same	LeakyReLU
	Conv2D	256	(1,3)	(1,2)	Same	LeakyReLU
Spatial analysis	Conv2D	1024	(Input size,1)	(1,1)	Valid	LeakyReLU
	Conv2DTrans	256	(Output size,1)	(1,1)	Valid	LeakyReLU
	Conv2DTrans	256	(1,3)	(1,2)	Same	LeakyReLU
	Conv2DTrans	128	(1,3)	(1,2)	Same	LeakyReLU
Temporal decoder	Conv2DTrans	64	(1,3)	(1,2)	Same	LeakyReLU
	Conv2DTrans	32	(1,3)	(1,2)	Same	None
	Conv2D	1	(1,1)	(1,1)	Valid	None

Conv2D = convolutional 2D layer, Conv2DTrans = deconvolutional 2D layer, LeakyReLU = leaky rectifying linear layer. For filter size and strides, the first number is for the spatial dimension and the second for the temporal dimension. In the spatial section, 'Input' is the number of input electrodes to the network and 'Output' is the number of output electrodes from the network. Padding represents if zero-padding is applied during convolutions, where 'Same' indicates that it is used and 'Valid' that it is not used.

Fig. 2. Examples of network generated artifacts. Original signals are in red and network generated are in blue. The network had six layers and a stride of two. The upper signal shows a periodic transient with spiky morphology at a frequency of 2 Hz. The lower signal also shows a periodic artifact, here at 8 Hz, and high frequency content of low amplitude. The generated signals seem to cyclically return to zero, most apparent for higher amplitudes.

chosen configuration of four layers and a stride of two most consistently produced results deemed by the specialist as most realistic. Using more layers tended to yield high frequency or periodic artifacts (Fig. 2).

This may be due to the deconvolutions and the 'checkerboard' problem (Odena et al., 2016). Inaccurate phases of the waveforms were not seen frequently but more often when using fewer layers or a stride of one. The tested filter sizes gave comparable results. From a practical point of view, using a stride of one and more layers produced larger networks, resulting in longer processing times and slower learning. For example, a network with two layers and a stride of two took 30 s per epoch to train and the final validation loss was reached after the analysis of a couple of epochs. This can be compared to a network with seven layers and a stride of one which consumed around 50 min per epoch and learned slowly—in this case training was discontinued since its completing would have consumed several weeks. Our final choice optimized the balance between visual results and processing times.

Training with mini batches, up to a size of 128, did not improve results; this was tested with and without batch normalization. Dropout layers did not improve results. Leaky rectifying linear units performed better than rectifying linear units; an alpha of 0.2 was used.

Optimization was performed with Adam (Kingma and Ba, 2015); no

other algorithm was tested. Learning rates above 10^{-4} resulted in instability during training, with sudden increases in the loss and deterioration of the appearance of the generated signals. So, for the final run this rate was chosen. Using a linear decay for the learning rate did not improve results. Decay parameters $\beta_1 = 0.5$ and $\beta_2 = 0.99$ were used. Mean absolute error was used as the loss function. Least square error was tested as well as versions of mean absolute and least square error with an electrode gradient that gave a larger penalty for errors of electrodes with consistently higher errors relative to other electrodes. All weights of the convolutional and deconvolutional layers were initialized according to the default setting, i.e., the kernel as a Glorot uniform distribution and the bias set to zero. No further regularization was used in the networks.

2.3.3. Training schedule

Training was performed with overlapping examples, where an example could start at any sample point within a recording allowing extraction of a 10 s example. An epoch was defined as training corresponding to one example from all (1,114) subjects one time. For the subsets, this meant training 223, 74, 37, 22, 11 and 2 times with each subject per epoch.

The training order of the subjects was randomized for each epoch.

For subjects with several EEG recordings, one was chosen at random. A start position defining a 10 s example was then randomly selected. If a drawn example from a subject was discarded due to high amplitude, attempts were made to redraw a new example with acceptable amplitude up to 100 times. If the subject had several recordings and the number of attempts was exceeded, then the next recording was tried and so on. If the number of attempts was exceeded for all recordings, no training took place for that subject during that epoch.

When training the GN3 network, 1 of the 21 input electrodes was randomly selected and the input was replaced by noise of normal distribution with mean 0 and a standard deviation of 0.1.

To decide the number of epochs, all networks were trained for up to 1,000 epochs, the epoch of minimum loss for the validation data was identified (Table 3), which was then used for the final run. For sessions with subject count below 50, a deterioration of performance was seen after some interval of training but in the other cases a stagnation of improvement was seen at some point but with no sign of overfitting.

One epoch of training usually took 40–60 s, depending on GPU and network. Evaluation during training to monitor overfitting was done at the end of each epoch with 1,000 examples from training and validation data, respectively, which added 30–60 s. Total training time for 1,000 epochs with evaluation at each epoch could be 15–30 hours.

2.4. Spherical spline interpolation

As a baseline for the evaluation of the networks, spherical spline interpolation was used to recreate the same signals of the test data set as the convolutional networks. This method was chosen since it seems to be one of the best performing of the standard interpolation techniques (Perrin et al., 1987, 1989; Soong et al., 1993) and is used in several common EEG processing softwares. The method consists of fitting a spherical surface to the known values of some electrodes and their corresponding spatial positions. Interpolated values can then be extracted from the resulting deformation of the surface (Freedman, 1984).

2.5. Evaluation

5,000 randomly chosen examples of the test data set were used for the final evaluation. The networks were evaluated in several ways using different methods and measures. The measures were: mean absolute error (MAE), correlation coefficient (R), coherence (C), cross spectral phase (P), Fréchet distance (FD), and Kullback-Liebler divergence (KL). They all reflect some aspect of the closeness or distance between generated and original data. Before calculating the measures, all data were renormalized to reflect real amplitudes (μV).

MAE was calculated straightforwardly by taking the absolute value of the difference between the original and recreated value x_i and y_i at each data point i and then taking the mean of all values N as $MAE = \frac{1}{N} \sum_{i=1}^N |y_i - x_i|$.

R was calculated between the original and recreated signals with the Spearman's rank correlation coefficient. Spearman's method was used since the relatively high occurrence of artifacts and biological signal transients resulted in a non-Gaussian distribution.

The coherence C was calculated as $C = \frac{|C_{xy}(f)|^2}{C_{xx}(f)C_{yy}(f)}$, where $C_{xy}(f)$ is the Fourier transform of the cross-correlation, $C_{xx}(f)$ and $C_{yy}(f)$ are the Fourier transforms of the auto-covariances of the signals x and y . The

Table 3

The number of epochs of training for the networks GN1, GN2 and GN3 when evaluating the effect of number of subjects in training data.

Number of subjects	5	15	30	50	100	500	1,114
GN1	60	100	100	400	1,000	1,000	1,000
GN2	100	100	100	400	400	1,000	1,000
GN3	100	100	100	400	400	1,000	1,000

phase P was calculated as the angle between the real and imaginary part of $C_{xy}(f)$. Values for frequencies up to 40 Hz were used.

FD is a measure of distance between two curves which also reflects differences in shape and was calculated using recursive algorithms (Eiter and Mannila, 1994). The analysis was limited to 500 examples of randomly chosen 2 s intervals due to the long processing times of the recursive algorithm. Processing one electrode consumed 8–10 s for 2 s intervals, which can be compared to 3–5 minutes for 10 s intervals. For the GN1 network with 17 electrodes and processing 500 examples, this corresponds to approximately 19–24 hours and 18–30 days, respectively.

KL is an entropy-based measure of the difference between the distributions P and Q of two data sets (Kullback and Liebler, 1951), where P in this case were the original data and Q the recreated data. It is calculated as $KL(P \parallel Q) = \sum p_i(x) \log \left(\frac{p_i(x)}{q_i(x)} \right)$. The distributions were calculated by binning the data in the interval $(-500, 500) \mu\text{V}$ using bins of size $5 \mu\text{V}$.

2.6. Comparison with spherical spline interpolation

Signals recreated by the networks were compared to the corresponding signals recreated by spherical spline interpolation. MAE, R, C, P, FD, and KL were calculated for each paired data point of original and recreated signals. Wilcoxon signed rank test was used to assess differences in the median value of the measures; for P the absolute value was used.

Network prediction and interpolation processing time were measured for 5,000 examples, using a computer with an Intel Xeon E5 1620V4 Quad Core 3.5 GHz processor, 96 GB 2,400 MHz RAM, and Nvidia Quadro P5000 GPU.

2.7. Visual assessment of data

Blinded tests were performed where five senior consultants in clinical neurophysiology (from our clinic, not otherwise involved in this project) visually assessed randomly displayed real, network generated (GN1) or interpolated EEG data. Their task was to determine whether each presented example was real, or computer generated. The purpose was to validate the visual quality, i.e., that wave morphology and fields appeared credible.

125 of the 5,000 test examples were chosen at random. The 125 examples were in turn randomized to either being displayed to the raters as the original signals, the signals recreated by the GN1 network, or by interpolation. The same randomization was used for all raters. A Python script was developed which displayed one example at a time. For each example, the rater had to press one of two buttons in the GUI to rate it as real or artificial. Pressing a button automatically loaded the next example with no possibility of going back. The first 25 examples were used for practice, after which it was verified that the rater had understood the task, and the remaining 100 examples were then used for the test. Each neurophysiologist had 20 min to complete the test part. The data was displayed in average montage, the amplitude could be altered but no other adjustments were possible.

A Chi-squared test was used to assess differences. The median rating of each example was used, and testing was done pairwise between data types.

2.8. The effect of the number of subjects

The effect of the number of subjects, also indirectly reflecting the amount of data, used for training were evaluated with MAE, R and C. A Kruskal-Wallis H test was used for multiple comparisons and when the null hypothesis could be rejected, post hoc analyses were performed by Dunn's multiple comparison test using an adjusted p-value

corresponding to 0.05.

3. Results

3.1. General

To give an impression of the signal quality and visually exemplify the significance of MAE, examples of data of MAE in the range 0–10 μV are given in Fig. 3. Each signal was randomly chosen from a randomly chosen EEG, with the exception that drawn examples containing no obvious biological signal (e.g., square waveform signal) were discarded and a new example was drawn.

There seemed to be a correlation for larger errors and increasing amplitude of the original signal. Comparing two electrodes illustrates that for similar MAE values the distributions of the errors may be different—smaller errors of uniform distribution vs. larger local errors: The signal of the Fp1 electrode at 9 μV encompassed blink and muscle artifacts. The blink artifacts were reproduced, whereas the muscle artifacts were not. The latter hence mainly contributed to the error. For the F8 electrode at 9.5 μV , the original and recreated signals follow each other fairly well, but in the interval 9.5–10.0 s, the original signal demonstrated an artifact of higher amplitude that is not reproduced.

The electrodes' relative positions on the scalp surface were approximated by a 2-dimensional grid geometry with their respective median value of the MAE (Fig. 4). For the GN1 network, a gradient was seen with increasing error in central to peripheral direction, which was also seen for the GN3 network, although not as prominent. There was also a gradient with increasing errors in posterior to anterior direction, which was seen in all networks.

3.2. Comparison with spherical spline interpolations

For the GN1 network, the 'constrained optimization by linear approximation' algorithm used in the MNE package usually did not converge. This was expected since the problem of interpolating 17 values from 4, where in addition some of the values to be interpolated are spatially distant, is underdetermined.

Overall, most measures showed a better performance for the convolutional networks compared to the spherical spline interpolation; 15 out of the 18 calculated values were significantly better (Fig. 5). However, the absolute difference between recreating the signals with convolutional networks or by interpolation were relatively small.

MAE, FD and R were all in favor of the networks. For KL the divergence was smaller for the GN1 and GN2 but larger for the GN3. C and P were better for GN2 and GN3, with higher coherence and less difference in phase compared to interpolation, but the opposite applied to GN1.

Performing predictions using a GPU gave processing times that were shorter than for interpolation (using CPU); Table 4. Predicting with CPU increased the processing times by a factor of 3–11, depending on network used, and consumed more time compared to interpolation.

3.3. Visual tests

The randomization gave a distribution of 34, 33 and 33 examples of original, network generated and spline interpolated data with 88, 82 and 52 percent rated as real (Fig. 6), respectively. There was no difference between the real and network generated data, whereas both performed better when compared to interpolated data.

For the original data there were none, for the GN1 network two, and for the interpolation eight examples where all raters were in full agreement as to them being artificial. These examples were visually inspected by author MS (specialist in clinical neurophysiology), and the impression was that they looked cleaner than the originals, i.e., fewer artifacts, and less variation resulting in already monomorphic signals looking even more monomorphic (Fig. 7). This was in line with spontaneous comments by some raters after the test, and some also thought that some examples looked 'too good'.

3.4. The effect of the number of subjects

An increase in performance was seen by using more subjects for training. The largest effect was seen for up to 100 subjects, after which only a small additive effect was seen (Fig. 8). However, the magnitude of the improvement created by adding subjects, depended on the network and parameter assessed.

3.5. (Post hoc) Comparison of data with normal, pathological, or muscle artifact content

To approach a possible clinical application where normal signals should be differentiated from various kinds of pathology, we visually assessed 1,000 of the 5,000 test examples for the GN1 network. The original signals were visualized and data with one of three characteristics were extracted: normal EEG, pathological EEG, and electrodes

Fig. 3. Examples of recreated EEG signals with different MAE to illustrate the visual significance of this measure. Signals were randomly selected with MAE spanning the interval 0–10 μV . The level of the baseline of each signal relative to the vertical axis corresponds to its MAE value. The original signal is in red and the signal recreated by GN1 is in blue.

Time in seconds (s) for evaluating 5,000 examples with predictions by the networks GN1, GN2 and GN3 or by spherical spline interpolation for the corresponding problems.

Stacked bar chart showing the percentage of Artificial (red) and Real (blue) data for three methods: Original, Network, and Interpolation. The y-axis represents the percentage from 0% to 100%. The x-axis labels are Original, Network, and Interpolation. A legend indicates Red for Artificial and Blue for Real. P-values are shown above the bars: $p = 0.001$ for the overall comparison, $p = 0.461$ for Original vs Network, and $p = 0.009$ for Network vs Interpolation.

Method	Real (%)	Artificial (%)
Original	~88%	~12%
Network	~82%	~18%
Interpolation	~52%	~48%

containing muscle artifacts. Normal data featured a clear posterior dominant rhythm in the alpha band. Pathological data contained increased low frequency content and asymmetries, and some had epileptiform activity. Only examples with very few artifacts were selected for normal and pathological data and all electrodes were used

For artifacts, after adjusting for multiple comparisons (Bonferroni), there were no significant differences between the network and interpolation (Fig. 10). Compared to normal and pathological data, the errors were larger and the correlation weaker.

We showed that the convolutional neural networks could recreate EEG signals with a higher precision and produce signals with a more credible visual appearance than one of the most commonly used techniques, spherical spline interpolation. Previous studies (Corley and Huang, 2019; Kwon et al., 2019) have indicated a performance superior to different interpolation methods using smaller data amounts, but comparisons are difficult due to large methodological discrepancies. In

Fig. 7. Example of interpolated data (blue) and original signals (red), which all raters assessed as artificial.

Fig. 8. MAE, correlation and coherence as a function of the number of subjects (5, 15, 30, 50, 100, 500, 1,114) used for training of the networks GN1 (blue), GN2 (red) and GN3 (green). Solid lines are the median, dashed lines the 25th and 75th percentiles, and the filled circles mark the points up to which there is a significant effect of adding more subjects. (Kruskal-Wallis H test were used for multiple comparisons and post hoc analyses were performed by Dunn's multiple comparison test using an adjusted p-value corresponding to 0.05.).

Fig. 9. Distributions of normal, pathological, and muscle artifact data (histograms normalized to one). The data ranges were -106 – 74, -213 – 271, and -195 – 181 μ V, respectively. The distributions were non-gaussian when tested for normality (Kolmogorov-Smirnov and Shapiro-Wilk).

the work by Corley et al. (2018), the units of MAE values are not given. Kwon et al. (2019) shows high correlations (up to 0.99) but their evaluation data likely have a better signal-to-noise level since it consisted of either simulated or trial averaged ERP data. Here we used randomly selected examples from a large pool of unsorted real data.

In this work, the focus was not on testing a large variety of architectures and parameters. We used a fairly simple and straightforward network architecture without extensive finetuning and still outperformed the standard method. This means that there are probably better network architectures and hyperparameters so that the performance might be further improved.

Whether the processing times are perceived as acceptable depends, apart from the network used, on the available hardware and the amount

of data to be interpolated. Using CPU for a standard recording of 20 min consumed 5.6–11.4 s for prediction, with the computer used for the tests. That would to many be seen as an acceptable duration. This can be compared to 1.0–2.7 s for interpolation, which of course is significantly shorter, but if a GPU is available the processing times for predictions will be slightly shorter, 0.7–2.0 s. Another issue is the time it may take for a new user to setup the method. For a user already having a computer running a Linux operating system with installed (Nvidia) graphics card and drivers, following the instructions given at our GitHub repository for setting up an Anaconda environment for a demo for the GN1 type network may be accomplished in fifteen minutes. However, given the large variability in different platforms and dependencies, the setup may take more time. In future work we will look into providing a Docker

Fig. 10. Comparisons between network generated (blue) and interpolated (red) data. Box-and-whiskers plots (minimum, maximum, median, and first and third quartiles). N = normal data ($n = 340$ electrodes), P = pathological data ($n = 544$ electrodes), and A = muscle artifacts ($n = 144$ electrodes). Left: MAE in μV . Middle: MAE calculated from signals normalized to a standard deviation (std) of one. Right: Correlation. P-values (not adjusted for multiple comparisons) are from Wilcoxon signed-rank test.

container that can be used to process any EEG dataset stored in a standardized format (e.g., EEG-BIDS) on any platform (Gorgolewski et al., 2017; Permet et al., 2019).

The amount of data used for our study was relatively large and the networks were trained with an even distribution of examples with respect to the number of subjects. The largest increase in performance was seen in the range 5–100 subjects, with only slight increases thereafter. This corresponded to a maximum of 111–278 hours of training data, a number that might be reduced given the overlapping training schedule. The impact of the number of subjects vs. the amount of data was not investigated but it is reasonable to assume that the amount of data per subject also influences the results, including the number of subjects needed to saturate the improvement. Generally, short recordings tend to yield small-, long recordings large variability. If a data set used for training has an unequal distribution in terms of recording length per subject it may be beneficial to prioritize training evenly with respect to the number of subjects. This is to capture the interindividual variations and avoid bias toward subjects with more data. However, very short recordings may give a low variation with the risk of overfitting and inferior generalization of the networks. The distributions of normal vs. pathological and of physiological variations in the data as well as the occurrence of artifacts, were largely unknown and not controlled for in the analysis. As judged from the material of the post hoc analysis, the signals used in this study contained relatively frequent artifacts and pathological features. Using data with well described distributions and balanced data sets, may improve performance further. However, assessing a data material of the size used here in such detail, is an extensive undertaking.

Even though the dataset used in this work was large, this does not guarantee that the results generalize well to other datasets. If the networks trained here are applied to, or if new networks are trained using data from a different database, the performance and the number of subjects needed to saturate the increase in performance may differ compared to our results.

Evaluating how different types of artifacts affect the results, and how wave transients and spatially restricted activity are reproduced is important. A limited evaluation of data of different character was made in the post hoc analysis, but more extensive testing needs to be performed. It is reasonable to assume that signals are more likely to be reproduced correctly if they are synchronized and have a wider spatial distribution, with the opposite being true for localized and random signals, since the prediction of one signal from another depend on the existence of information of the first signal in the second (Lauritzen, 1974). Since synchronization is a hallmark of cortical activity (Nunez and Srinivasan, 2006a), the prediction of the activity in one location from another is plausible. However, since small variations in timing have a higher probability of causing destructive interference for higher frequencies, maintaining phase-synchronization over longer distances may be harder for these frequencies. Moreover, there is evidence that higher frequencies have a more limited spatial synchronization due to resonance phenomenon (Nunez and Srinivasan, 2006b). This could

mean that predicting high frequency activity at low electrode densities may be harder.

In the post hoc analysis, data of normal and pathological character was compared. The pathological data often had asymmetries, components of lower frequencies, and focal fast transients (e.g., epileptiform activity). The network predictions still were more accurate compared to interpolation but the difference in performance was small. For both methods, the absolute errors were larger for the pathological data, but when normalizing by the standard deviation, the errors were comparable for the respective data type. We interpreted this result as the pathological data on average having larger amplitudes and the errors simply being roughly proportional to the amplitude, rather than the larger errors being due to inferior recreation of asymmetries or other characteristics of the data type.

There are many types of possible artifacts that may be superimposed on the cortical signals and these usually have certain characteristics with respect to morphology, spatial distribution, synchronization and randomness (Tatum et al., 2018). One common source of artifacts is muscle activity. This generates a relatively stochastic electrical signal (Reaz et al., 2006), resulting in an artifact with a random character and high frequency content. In the post hoc tests, there was no significant difference between network prediction and interpolation for muscle artifacts. The errors were larger compared to the other data types and the correlations were weaker. An example of muscle artifacts, not recreated by the network and hence contributing to the error, is seen in Fig. 3 at a MAE of 9 μV . The signals from frontal and anterior temporal electrodes contained more muscle artifacts than those of other electrodes. This agreed with the results of the post hoc analysis where most of the signals containing muscle artifacts were found in the signals from these electrodes. Possibly in line with this, a posterior to anterior gradient was noted, with larger errors in the frontal quadrants (Fig. 3). Eye movements, producing a low-frequency, more organized artifact, affect frontal electrodes the most. In Fig. 3, blink artifacts were indeed reproduced. It is possible that the reproducibility may vary depending on variation of the spatial distribution of the artifact. Blink artifacts usually are of relatively large amplitudes, the post hoc analysis indicates that the absolute error was larger for larger signal amplitudes, and this may suggest that these artifacts contribute to larger errors and thus also contribute to the error gradient found. Probably adding further to this error gradient is the higher degree of synchronization over posterior parts due to posterior dominant rhythms. There was also an increasing error in radial direction (Fig. 4). This is probably an effect of distance, since predicting over longer distances is a harder problem. For artifacts, it may not be desirable that the method recreates the signal exactly. The interpretation of the measures may reverse, e.g., a higher MAE may reflect a better performance if a blink artifact is not reproduced. However, since there usually is no ground truth for the cortical signal, it is impossible to know if it is recreated correctly.

In addition to the further need to evaluate how the networks recreates various waveforms, e.g., epileptiform activity, the clinical usefulness must be evaluated. Three-way comparisons of clinical diagnosing

using i) a full EEG montage recreated by a network, ii) the reduced EEG montage from which the network upsamples, and iii) a full EEG montage of original data as baseline need to be performed. The same problems as those of assessing the sensitivity of reduced montages, identified by Westover et al. (2020), apply to assessing the sensitivity of the networks: electrode positions vs. EEG patterns, interrater variability, and asymmetric comparisons between reduced and full montages. In line with previous discussions in this section and as implied by earlier studies (Bennis et al., 2017; Foldvary et al., 2000; Kolls and Husain, 2007), reduced montages probably have lower sensitivity for focal activity. For patterns with general distributions, e.g., in the setting of anoxic brain injury in cardiac arrest with highly malignant patterns (Backman et al., 2020), studies show high sensitivity for reduced montages. Hence, upsampling may be more useful for focal compared to general patterns. However, in the clinical setting, the disease course is dynamic, and it is not *a priori* known what patterns will emerge. Detecting patterns early on that could indicate higher risk of seizure or status epilepticus, e.g., focal epileptiform activity, may benefit from more electrodes. The sensitivity may also depend on the EEG interpreters' training and experience. Possibly, an increase in sensitivity with more electrodes is more likely for inexperienced interpreters. When comparing a network with a reduced montage, comparisons must be made symmetrically, i.e., the raters should have access to the same amount of data and ancillary information. In summary, testing should reflect clinical conditions for interpretation and be made with attention to the problems outlined by Westover et al. (2020).

In this work, we showed that networks can upsample and restore signals. This suggests that the networks are robust to noise or missing information, which have an impact on how they perform during detection or classifications tasks. It is not known if the networks learned the statistical distributions of the electrical fields or if they just learned a simpler way of interpolating the values. The visual tests suggest that the signals recreated by the network have a more natural variation, which may imply that the former is the case. Ordinary interpolation does not add information, so upsampling data by interpolation should not increase the accuracy of neural networks. Adding a general knowledge of the statistical distributions could theoretically boost performance. Thus, training networks having a low number of input channels to also learn the distributions of a larger set of channels might accomplish such a performance boost.

The clinical usefulness of the kind of networks developed in this study remains to be proven. The performance improvement was relatively small compared to interpolation and the processing times (with CPU) were longer. However, even though the difference in performance was small, our impression was that the data generated by the networks had a more 'life-like' quality compared to interpolation, as indicated by the visual test. Moreover, the networks can also be trained to perform multiple different processing, a limited demonstration of this was provided by the GN3 network which both detected missing electrodes and recreated the signals. It may well be possible to develop networks that both process EEG data (e.g., upsample, restore bad electrodes, and remove artifacts), as well as performing classification tasks (e.g., sleep staging, detecting epileptiform activity or seizures). Hence, compared to interpolation, convolutional neural networks seem to have more potential for improving the performance further. They may also, despite constituting a complex method, be a more efficient practical implementation compared to an elaborate EEG processing pipeline consisting of many different methods for each type of processing step.

5. Conclusions

We demonstrated that restoring or interpolating EEG signals can be performed with neural networks with a precision as good as, or better than, spherical spline interpolation and often with a more credible appearance. Processing times were acceptable for standard EEG exams. It may be worth-while using up to 100 subjects for training a network for

this type of task. Minor further gains in performance were possible by adding more subjects but may not be worth the effort if data is hard to acquire. There may be room for improvement by testing other network architectures and using a more content-controlled data. The present networks need further evaluation to assess how they reproduce different important wave phenomena and spatial asymmetries, and how different artifacts affects the results. The method could potentially be used clinically to recreate full montage EEGs from long term monitoring using four electrodes and recreate low quality channels to provide better visualization. The limited increase of performance and its complexity compared to spherical splines interpolation makes it probable that the method needs further development. The usefulness of such an application has to be evaluated clinically. However, on the whole the results were promising.

Author contributions

Mats Svantesson: Conceptualization; Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Visualization; Writing - original draft; Writing - review & editing. Håkan Olausson: Writing - review & editing. Anders Eklund: Resources, Supervision; Writing - review & editing. Magnus Thordstein: Writing - review & editing.

Declaration of Competing Interest

None.

Acknowledgements

The work was mainly funded by Linköping University, the University Hospital of Linköping, and the ALF of Region Östergötland (LIO-936176 and RÖ-941359). AE was supported by the ITEA3/VINNOVA funded project Intelligence based iMprovement of Personalized treatment And Clinical workflow support (IMPACT). We would also like to thank the Nvidia corporation, who donated the Quadro P6000 and Titan X graphics cards used for training the networks.

References

- Acharya, U.R., Sree, S.V., Swapna, G., Joy Martis, R., Suri, J.S., 2013. Automated EEG analysis of epilepsy: a review. *Knowledge Based Syst.* 45, 147–165. <https://doi.org/10.1016/j.knsys.2013.02.014>.
- Aznan, N.K.N., Atapour-Abarghouei, A., Bonner, S., Connolly, J.D., Al Moubayed, N., Breckon, T., 2019. Simulating brain signals: creating synthetic EEG data via neural-based generative models for improved SSVEP classification. *International Joint Conference on Neural Networks (IJCNN)*.
- Backman, S., Cronberg, T., Rosén, I., Westhall, E., 2020. Reduced EEG montage has a high accuracy in the post cardiac arrest setting. *Clin. Neurophysiol.* 131 (9), 2216–2223. <https://doi.org/10.1016/j.clinph.2020.06.021>.
- Bennis, F.C., Geertsema, E.E., Velis, D.N., Reus, E.E., Visser, G.H., 2017. The use of single bipolar scalp derivation for the detection of ictal events during long-term EEG monitoring. *Epileptic Disord.* 19 (3), 307–314. <https://doi.org/10.1684/epd.2017.0926>.
- Corley, L., Huang, Y., 2019. Deep EEG super-resolution: upsampling EEG spatial resolution with generative adversarial networks. Las Vegas, Nevada, USA, IEEE EMBS International Conference on Biomedical & Health Informatics 100–103.
- Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *J. Neurosci. Methods* 134, 9–21. <https://doi.org/10.1016/j.jneumeth.2003.10.009>.
- Eiter, T., Mannila, H., 1994. Computing Discrete Fréchet Distance. Vienna, Austria: Tech. Report CD-TR 94/64 Christian Doppler Laboratory for Expert Systems.
- Fletcher, E.M., Kussmaul, C.L., Mangun, G.R., 1996. Estimation of interpolation errors in scalp topographic mapping. *Electroenceph. Clin. Neurophysiol.* 98, 422–434. [https://doi.org/10.1016/0013-4694\(96\)95135-4](https://doi.org/10.1016/0013-4694(96)95135-4).
- Foldvary, N., Caruso, A.C., Mascha, E., Perry, M., Klem, G., McCarthy, V., Qureshi, F., Dinner, D., 2000. Identifying montages that best detect electrographic seizure activity during polysomnography. *Sleep* 23 (2), 221–229.
- Freedman, W., 1984. Spherical spline interpolation-basic theory and computational aspects. *J. Comput. Appl. Math.* 11, 367–375. [https://doi.org/10.1016/0377-0427\(84\)90011-6](https://doi.org/10.1016/0377-0427(84)90011-6).
- Friberg, H., Westhall, E., Rosén, I., Rundgren, M., Nielsen, N., Cronberg, T., 2013. Clinical review: continuous and simplified electroencephalography to monitor brain

- recovery after cardiac arrest. *Crit. Care* 17 (4), 233. <https://doi.org/10.1186/cc12699>.
- Gorgolewski, K.J., Alfaro-Almagro, F., Auer, T., Bellec, P., Capotà, M., Chakravarty, M. M., Churchill, N.W., Cohen, A.L., Craddock, R.C., Devenyi, G.A., Eklund, A., Esteban, O., Flandin, G., Ghosh, S.S., Guntupalli, J.S., Jenkinson, M., Keshavan, A., Kiar, G., Liem, F., Raamana, P.R., Raffelt, D., Steele, C.J., Quirion, P.O., Smith, R.E., Strother, S.C., Varoquaux, G., Wang, Y., Yarkoni, T., Poldrack, R.A., 2017. BIDS apps: improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS Comput. Biol.* 13 (3), e1005209 <https://doi.org/10.1371/journal.pcbi.1005209>. PMID: 28278228; PMCID: PMC5363996.
- Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., Hämäläinen, M., 2013. MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* 7 <https://doi.org/10.3389/fnins.2013.00267>. ISSN 1662-453X.
- Gururangan, K., Razavi, B., Parvizi, J., 2018. Diagnostic utility of eight-channel EEG for detecting generalized or hemispheric seizures and rhythmic periodic patterns. *Clin. Neurophysiol. Pract.* 3, 65–73. <https://doi.org/10.1016/j.cnp.2018.03.001>.
- Hartmann, K.G., Schirrmeyer, R.T., Ball, T., 2018. EEG-GAN: generative adversarial networks for electroencephalographic (EEG) brain signals. *arXiv preprint arXiv:1806.01875*.
- Hera, J., Koren, J., Fürbass, F., Hartmann, M., Gruber, A., Baumgartner, C., 2017. Reduced electrode arrays for the automated detection of rhythmic and periodic patterns in the intensive care unit: Frequently tried, frequently failed? *Clin. Neurophysiol.* 128, 1524–1531. <https://doi.org/10.1016/j.clinph.2017.04.012>.
- Jasper, H.H., 1958. The ten twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.* 10, 367–380.
- Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. *San Diego, International Conference on Learning Representations*.
- Koles, Z.J., Paranjape, R.B., 1988. Topographic mapping of the EEG: an examination of accuracy and precision. *Brain Topogr.* 1 (2), 87–95. <https://doi.org/10.1007/BF01129173>.
- Kolls, B.J., Husain, A.M., 2007. Assessment of hairline EEG as a screening tool for nonconvulsive status epilepticus. *Epilepsia* 48 (5), 959–965. <https://doi.org/10.1111/j.1528-1167.2007.01078.x>.
- Kubota, Y., Nakamoto, H., Egawa, S., Kawamata, T., 2018. Continuous EEG monitoring in ICU. *J. Intens. Care* 6, 38. <https://doi.org/10.1186/s40560-018-0310-z>.
- Kullback, S., Liebler, R.A., 1951. On information and sufficiency. *Ann. Math. Stat.* 22, 79–86. <https://doi.org/10.1214/aoms/1177729694>.
- Kwon, M., Han, S., Kim, K., Jun, S.C., 2019. Super-resolution for improving EEG spatial resolution using deep learning convolutional neural network-Feasibility study. *Sensors* 19, 5317. <https://doi.org/10.3390/s19235317>.
- Lauritzen, S.L., 1974. Sufficiency, prediction and extreme models. *Scand. J. Stat.* 128–134.
- Luo, Y., Lu, B.L., 2018. EEG data augmentation for emotion recognition using a conditional Wasserstein GAN. *International Conference Engineering in Medicine and Biology Society* 2535–2538.
- Luo, T., Fan, Y., Chen, L., Guo, G., Zhou, C., 2020. EEG signal reconstruction using a generative adversarial network with Wasserstein Distance and temporal-spatial-Frequency loss. *Front. Neuroinform.* 15 <https://doi.org/10.3389/fninf.2020.00015>.
- Ma, B.B., Johnson, E.L., Ritzl, E.K., 2018. Sensitivity of a reduced EEG montage for seizure detection in the neurocritical care setting. *J. Clin. Neurophysiol.* 35 (3), 256–262. <https://doi.org/10.1097/WNP.0000000000000463>.
- Maas, H.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. *Atlanta, Georgia, USA, International Conference on Machine Learning*.
- Millan, J., 2004. On the need for on-line learning in brain-computer interfaces. In: *IEEE International Joint Conference on Neural Networks*, 4, pp. 2877–2882. <https://doi.org/10.1109/IJCNN.2004.1381116>.
- Nahrstaedt, H., Lee-Messer, C., 2019. holgern/pyedflib 2017. URL: <https://github.com/holgern/pyedflib> (accessed 21 September 2019).
- Nunez, P.L., Srinivasan, R., 2006a. Human alpha rhythms. *Electric Fields of the Brain*. Oxford University Press, New York, pp. 12–19.
- Nunez, P.L., Srinivasan, R., 2006b. Neocortical dynamics, EEG and cognition. *Electrical Fields of the Brain*. Oxford University Press, New York, p. 510, 369, 401–402.
- Obeid, I., Picone, J., 2016. The temple university hospital EEG data corpus. *Front. Neurosci.* 10, 1–5. <https://doi.org/10.3389/fnins.2016.00196>.
- Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and checkerboard artifacts. *Distill* 2016. <https://doi.org/10.23915/distill.00003>. URL: <https://distill.pub/2016/deconv-checkerboard/>.
- Oostenveld, R., Fries, P., Maris, E., Schoffelen, J., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* <https://doi.org/10.1155/2011/156869>. Article ID 156869.
- Pascual, D., Aminifar, A., Atienza, D., Ryvlin, P., Wattenhofer, R., 2019. Synthetic Epileptic Brain Activities Using GANs. *Machine Learning for Health, Vancouver, Canada*.
- Pati, S., McClain, L., Moura, L., Fan, Y., Westover, M.B., 2017. Accuracy of limited-montage electroencephalography in monitoring postanoxic comatose patients. *Clin. EEG Neurosci.* 48 (6), 422–427. <https://doi.org/10.1177/1550059417715389>.
- Pearl, P.L., Beal, J.C., Eisermann, M., Misra, S., Plouin, P., Moshe, S.L., Riviello, J.J., Nordli, D.R., Mizrahi, E.M., 2018. Normal EEG in wakefulness and sleep: preterm; Term; Infant; Adolescent. *Niedermeyer's Electroencephalography*. Oxford University Press, New York, pp. 167–171.
- Pernet, C.R., Appelhoff, S., Gorgolewski, K.J., Flandin, G., Phillips, C., Delorme, A., Oostenveld, R., 2019. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci. Data* 6 (1), 103. <https://doi.org/10.1038/s41597-019-0104-8>. PMID: 31239435; PMCID: PMC6592877.
- Perrin, F., Pernier, J., Bertrand, O., Giard, M., Echallier, J.F., 1987. Mapping scalp potentials by surface spline interpolation. *Electroenceph. Clin. Neurophysiol.* 66, 75–81. [https://doi.org/10.1016/0013-4694\(87\)90141-6](https://doi.org/10.1016/0013-4694(87)90141-6).
- Perrin, F., Pernier, J., Bertrand, O., Echallier, J.F., 1989. Spherical splines for scalp potential and current density mapping. *Electroenceph. Clin. Neurophysiol.* 72, 184–187. [https://doi.org/10.1016/0013-4694\(89\)90180-6](https://doi.org/10.1016/0013-4694(89)90180-6).
- Qiu, J., Zhao, W., 2018. Data encoding visualization based cognitive emotion recognition with AC-GAN applied for denoising. *IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing* 222–227. <https://doi.org/10.1109/ICCI-CC.2018.8482097>.
- Reaz, M.B.I., Hussain, M.S., Mohd-Yasin, F., 2006. Techniques of EMG signal analysis: detection, processing, classification and applications. *Biol. Proced. Online* 8 (1), 11–35. <https://doi.org/10.1251/bpo115>.
- Roy, Y., Banville, H., Alberquerque, I., Gramfort, A., Falk, T.H., Faubert, J., 2019. Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* 16 (5), 051001 <https://doi.org/10.1088/1741-2552/ab260c>.
- Satherley, B.L., Jones, R.D., Bones, P.J., 1996. EEG spectral topography in neurology: II. A new system and a theoretical comparison of interpolation techniques. *Australas. Phys. Eng. Sci. Med.* 19 (3), 183–193.
- Schomer, D.L., Epstein, C.M., Herman, S., Maus, D., Fisch, B.J., 2018. Recording principles: analog and digital principles; Polarity and field determinations; Multimodal monitoring; Polygraphy (EOG, EMG, ECG, SAO₂). *Niedermeyer's Electroencephalography*. Oxford University Press, New York, p. 134.
- Shapkin, A., 2012. Reading and Saving of Data in the EDF+. URL: <https://se.mathworks.com/matlabcentral/fileexchange/38641-reading-and-saving-of-data-in-the-edf> (accessed 1 August 2019).
- Soong, A.C.K., Lind, J.C., Shaw, G.R., Koles, Z.J., 1993. Systematic comparisons of interpolation techniques in topographic brain mapping. *Electroenceph. Clin. Neurophysiol.* 87 [https://doi.org/10.1016/0013-4694\(93\)90018-Q](https://doi.org/10.1016/0013-4694(93)90018-Q), 1985–195.
- Sun, C., Shrivastava, A., Singh, S., Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era. *IEEE International Conference on Computer Vision (ICCV)*. <https://doi.org/10.1109/ICCV.2017.97>.
- Tatum, W.O., Reinsberger, C., Dworetzky, B.A., 2018. Artifacts of recording and common errors in interpretation. *Niedermeyer's Electroencephalography*. Oxford University Press, New York, pp. 267–279.
- Tian, G., Qin, K., Wu, Y.M., Ji, Z., Wang, J.X., Pan, S.Y., 2012. Outcome prediction by amplitude-integrated EEG in adults with hypoxic ischemic encephalopathy. *Clin. Neurol. Neurosurg.* 114 (6), 585–589. <https://doi.org/10.1016/j.clineuro.2011.12.011>.
- Vanherpe, P., Schrooten, M., 2017. Minimal EEG montage with high yield for the detection of status epilepticus in the setting of postanoxic brain damage. *Acta Neurol. Belg.* 117 (1), 145–152. <https://doi.org/10.1007/s13760-016-0663-9>.
- Völker, M., Schirrmeyer, R.T., Fiederer, L.D.J., Burgard, W., Ball, T., 2018. Deep transfer learning for error decoding from non-invasive EEG. In: *The 6th International Winter Conference on Brain-Computer Interface*. Gangwon, South Korea. <https://doi.org/10.1109/IWW-BICI.2018.8311491>.
- Westover, M.B., Gururangan, K., Markert, M.S., Blond, B.N., Lai, S., Benard, S., Bickel, S., Hirsch, L.J., Parvizi, J., 2020. Diagnostic value of electroencephalography with ten electrodes in critically ill patients. *Neurocrit. Care.* <https://doi.org/10.1007/s12028-019-00911-4>.
- Zhang, Q., Liu, Y., 2018. Improving brain computer interface performance by data augmentation with conditional Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1806.07108*.