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

* Corresponding author. Department of Clinical Neurophysiology, University Hospital of Linköping, 58185 Linköping; E-mail address: mats.svantesson@liu.se

ARTICLE INFO

Keywords:

Deep learning

Convolutional neural networks

Electroencephalography

Signal reconstruction

Spatial upsampling

Spherical spline interpolation

ABSTRACT

In clinical practice, EEGs are assessed visually and recordings with reduced number of electrodes or artefacts make the assessment more difficult. Present techniques for upsampling or restoring channels utilize different interpolation strategies by taking averages of neighboring electrodes or fitting surfaces to the electrodes. These techniques usually perform better for higher electrode densities and values interpolated at areas far from electrodes can be unreliable. Using a method that instead learns the statistical distribution of the cortical electrical fields and predicts values of missing electrodes may yield better results.

Generative networks based on convolutional layers were trained to upsample from 4 or 14 channels or restore single missing channels to recreate 21 channel EEGs. Roughly 5,144 hours of data from 1,385 subjects of the Temple University Hospital EEG data Corpus were used for training and evaluating the networks. The results were compared to interpolation by spherical splines and a visual evaluation by board certified clinical neurophysiologists was conducted. In addition, the effect on performance due to the number of subjects used for training was evaluated.

The generative networks performed significantly better overall compared to spherical spline interpolation. There was no difference between real and network generated data in the number of examples assessed as fake by experienced EEG interpreters. On the contrary, the number was significantly higher for data generated by interpolation. Network performance improved with increasing number of included subjects, with the greatest effect seen for 100.

Using neural networks to restore or upsample EEG signals is a viable alternative to existing methods.

1 Introduction

The assessment of EEG in the clinical setting thus far has been 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.

In clinical practice, cortical electrical fields are usually recorded non-invasively using scalp electrodes placed according to the international 10-20 system (Jasper, 1958). The magnitude of the EEG signals is on the order of 10 – 100 μV and the signal-to-noise level is often low (Schomer et al., 2018). It is thus common for various artefacts (e.g., from eye movements, muscle activity, electrodes with high impedance etc.) to make EEG interpretation harder.

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 gives a low spatial resolution, making the recordings harder to assess compared to standard EEGs since identification and classification of wave phenomenon often depend on spatial patterns.

In the field of clinical neurophysiology, publications with deep learning applied to EEG are seen at an increasing rate. Typical applications are detection of epileptiform discharge, seizure detection, sleep staging, brain computer interfacing, and data processing and generation, where convolutional neural networks (CNNs) are most common, used in 40 % of the papers (Roy et al., 2019). The CNN architecture has similarities to the organization of the visual cortex (Felleman & van Essen, 1991; Hubel & Wiesel, 1962), where filters in each layer extracts position invariant features

in the data, the structure allow for a sparse connectivity and parameter sharing which improves statistical efficiency, reduces memory requirements and by stacking layers a hierarchy going from low- to high-level features is formed (LeCun et al., 1999). Since the filters learn the features, the need for feature engineering is eliminated; an advantage compared to other machine learning methods (LeCun et al., 2015). On the downside, it is usually necessary to train the networks with large amounts of data (Bach et al., 2017).

To be able to predict one signal from another there must exist information of the first signal in the second (Lauritzen, 1974). Synchronization is a hallmark of cortical activity (Nunez & Srinivasan, 2006), making the prediction of the activity in one location from another plausible. Since small variations in timing have a higher probability of causing destructive interference, it may be harder to maintain phase-synchronization over distances for higher frequencies. There is evidence that higher frequencies have a more limited spatial synchronization due to resonance phenomenon (Nunez & Srinivasan, 2006). This could mean that predicting high frequency activity at low electrode densities may be harder.

There are many types of possible artefacts that may be superimposed on the cortical signals and these usually have certain characteristics (Tatum et al., 2018). Some artefacts have a stereotypic morphology and a spatial distribution, e.g., the blink artefacts. Other artefacts have a more widespread distribution spatially with high degree of synchronization, e.g., electrocardiac artefacts. There are more local artefacts, such as those from poor contact electrode. Muscle activity generates a relatively stochastic electrical signal (Reaz et al., 2006), resulting in an artefact with a random character. In summary, 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. This applies to both cortical signals and artefacts.

When training deep neural networks for EEG processing, having data from an adequate number of individuals may be important for generalization but this has not been thoroughly investigated.

Using 1,500 ms epochs of 128-channel EEG data from up to 30 subjects in a cognitive task, an improvement in classification accuracy was seen with increasing number of subjects, most notably more than 15 subjects (Völker et al., 2018). The amount of data used for training was not reported and no statistical testing of the effect was presented in the paper. If the data set 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. However, if some subjects only have very short recordings this might result in a poor variation in training data. A longer recording may have a larger variation, for instance, the probability of drowsiness or sleep may increase. Given the huge inter- and intraindividual variation, finding a training data balance may be a challenge.

The main methods for scalp potential interpolation are nearest-neighbor or splines (Fletcher et al., 1996). Nearest-neighbor uses interpolated values based on the closest electrodes values while the splines method fit planes to the values of the electrodes, thereby forming a deformed but smooth surface from which interpolated values can be extracted. Spline techniques seem to perform better than nearest-neighbor with a tendency for spherical splines to perform best (Perrin et al., 1987; Perrin et al., 1989; Soong et al., 1993). However, 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 has also been claimed that the electrode density used clinically, i.e., the international 10-20 system, is too low for interpolation (Soong et al., 1993).

Surveying the deep learning field for papers on interpolating or upsampling EEG one paper was found. Wasserstein generative adversarial networks have been used to upsample from 8 or 16 to 32 channels (Corley & Huang, 2018). The method showed a reduction of the mean absolute error by one to two orders of magnitude when comparing to bicubic interpolation. This result was achieved using the training dataset 'V' of Berlin Brain Computer Interface Competition III (Millan, 2004), consisting of 2,141 s of 32 channel EEG sampled at 512 Hz registered from 3

subjects. The data was split to sections of 512 samples with a stride of 32 samples and the resulting data at a 75-20-5 percentage ratio for training, validation and testing. These results are somewhat difficult to evaluate since the units of the errors were left out in the paper.

There is a clinical need for better methods of processing EEGs. Therefore, in this study we investigate if CNNs can upsample the electrode density or restore corrupt signals of EEG with a precision comparable to the standard methods used today. We also test to what extent training the networks with different number of subjects affects the results, since this information may be useful when planning or evaluating studies using CNNs.

2 Materials and methods

2.1 Software and hardware

The computers used the operative 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 & C., 2017) 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 i7-6800K Hexa Core 3.4 GHz, 96 GB 2,133 MHz RAM, Nvidia Quadro P6000 and Nvidia Titan X.

2.2 EEG data

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

The recording electrodes were positioned according to the international 10-20 system (Jasper, 1958). A full EEG here is 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 had more electrodes but only those listed here were used. The data was not re-referenced.

A total of 1,385 (♂/♀: 751/634) subjects with 11,163 (♂/♀: 5,648/5,515) recordings, corresponding to 5,144 hours, was extracted from the data set based on the criterion described below. Ages varied from 18 to 95 years with a mean age of 51±18 years. All original recordings were in European Data Format (EDF), unfiltered and sampled at 256 Hz. An arbitrary lower limit of recording length of 300 seconds was used to ensure that each recording would offer some variance during training. There were some inconsistencies in the data set; for some subject folders there seemed to be different subjects in the same folder (e.g., female of 53 years and male of 21 years with a report for the female). For some, the birth year and date of recording in the header were irrational (e.g., born or recorded in the year 1899). Therefore, to be included, folders with multiple recording had to have matching gender and matching year or age for all EDF files and that the available information could be assumed to originate from an 18 years or older individual.

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 1,138, 131 and 143 subjects for each data set.

Data examples were generated randomly during training, validation and testing, see section 2.3.3; to avoid very large artefacts, drawn examples with maximum absolute amplitude over 500 µV were discarded. The first and last 40 s of each recording was also discarded since these parts sometimes had more artefacts or did not contain any cortical signal (usually a low frequency square wave). This time interval was arbitrarily chosen from visual sampling 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 artefacts.

Each example was normalized to a standard deviation of one. It was also tested to not normalize or to normalize each example to have a maximum amplitude of one, but the performance was inferior, with no normalization being the worst option, most notably a more pronounced tendency to underestimate the amplitudes. Since the data were high pass filtered, the mean was close to zero on average.

2.2.2 Ethical considerations

The HIPAA Privacy Rule was followed when the TUH EEG Corpus was created (Obeid & Picone, 2016). Since the used data was from a published data base with no subject information and

therefore no way of tracing the data back to the subjects, no further approval from an ethic committee was deemed necessary for this study.

2.3 The networks

2.3.1 Architecture

In this study, three generative networks were trained for slightly different tasks. The first (GN1) upsampled from 4 to 21 electrodes, the second (GN2) upsampled from 14 to 21 electrodes and the third (GN3) could recreate any of the 21 electrodes. The networks shared the same general architecture, the only essential difference was in the number of input and output channels. 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 (see Table 1). The function of the GN3 network differed since it was trained to dynamically replace anyone of the electrodes, meaning that it also had to learn to detect which channel it was supposed to recreate.

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 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, reducing each channel from an initial size of 2,560x1 while adding filters to a size of 160x256. Second, a spatial analysis and upsampling performed by convolution with a filter size corresponding to the number of input channels giving one channel with a size of 160x1,024, and then deconvoluting to the right number of channels, each having a size of 160x256. Third, a temporal decoder of four deconvolutional layers and a finishing convolutional layer to merge the filters resulting in the desired size 2,560x1 (see Table 2). GN1 and GN2 had 6,546,241 and GN3 had 6,808,385 parameters, all trainable.

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

Temporal decoder	Conv2DTrans	256	(1,3)	(1,2)	Same	LeakyReLU
	Conv2DTrans	128	(1,3)	(1,2)	Same	LeakyReLU
	Conv2DTrans	64	(1,3)	(1,2)	Same	LeakyReLU
	Conv2DTrans	32	(1,3)	(1,2)	Same	LeakyReLU
	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 channels to the network and 'Output' is the number of output channels from the network.

2.3.2 Hyperparameters

When testing different hyperparameters, a smaller set of recordings from the TUH EEG Corpus sampled at 250 Hz and consisting of 187 hours of data were used for training and validation (the length of each channel was adjusted to 2,500 samples). In the final run, the 256 Hz data set was used with training and validation subsets during training and the test subset after training. This may have seemed unnecessary, since the whole set only was used for the final training, but having the three sets was a safety measure taken in case further adjustments of the hyperparameters were required (which however there were not).

The four layers and a stride two were chosen empirically; it most consistently gave visually pleasing results and offered shorter training times compared to larger nets. The validation losses were usually about the same for nets with two to seven convolutional layers, but more layers showed a tendency for high frequency artefacts, and fewer layers to have inaccurate phases of waveforms. Inaccurate phases of waveforms were also seen when using a stride of one and while observing waveform generation during training. The formation of features was more evident for a stride of

two compared to one, with the stride of one giving an impression of just passing the signal through the network.

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 & Ba, 2015); no other algorithm was tested. Learning rates above 10^{-4} showed an instability during training, 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 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

When training, an example was chosen by first randomly selecting a subject. If there were several EEG recordings for the drawn subject, one was chosen at random. A start position defining a 10 s epoch was then randomly selected, hence resulting in an overlapping training schedule giving more variation in the training data compared to having fixed epochs. 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 were 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 channels was randomly selected and the input was replaced by noise of normal distribution with mean 0 and a standard deviation of 0.1.

The training data consisted of 1,114 subjects corresponding to 3,976 hours of data. To evaluate the effect of the number of subjects used, the 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 hours of data, respectively.

Since data examples could start at any sample point in a recording, there were many examples but also a large overlap between them. An epoch was therefore 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.

To decide the number of epochs, all networks were trained for up to 1,000 epochs and the loss as a function of epochs were assessed by visual inspection to identify the optimal number, defined as the lowest point before deterioration or stagnation of performance (Table 3). 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.

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

One epoch of training usually took 40 – 60 s, depending on GPU and network. Evaluation during training were often done 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 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; Perrin et al., 1989; Soong et al., 1993). 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 (Freeden, 1984).

2.5 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.

MAE was calculated straight forwardly 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 artefacts 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 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 & Mannila, 1994).

KL is an entropy-based measure of the difference between the distributions P and Q of two data sets (Kullback & 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) µV using bins of size 5 µV.

For each trained network 5,000 data examples were randomly selected from the 143 test subjects. With a duration of 10 s and a sample rate of 256 Hz the number of data points for the amplitude per example were 2,560 per channel. MAE was calculated for each example, thus giving 5,000 data points per channel. For the cross spectral density there were 41 data points per example for the assessed frequencies for values ≤ 40 Hz, giving a total of 205,000 data points each for C and P, per channel. For R and KL, 50 examples were used to calculate each value giving 100 data points per channel. For R, this was to attain certain values (very low p-values) and for KL it was to avoid having zero-values in the distributions causing divisions by zero. Calculating FD was very time consuming and therefore limited to 500 examples, giving as many data points, and only using 2 s of each example, where the time interval was chosen at random from each 10 s.

Table 1

A summary of the number of data points used for calculating the median for each measure and network.

	MAE	R	C	P	FD	KL
GN1	85,000	1,700	3,485,000	3,485,000	8,500	1,700
GN2	35,000	700	1,435,000	1,435,000	3,500	700
GN3	105,000	2,100	4,305,000	4,305,000	10,500	2,100

GN1, GN2 and GN3 have 17, 7 and 21 output channels, respectively. The measures MAE, C, R, P, FD and KL have 5,000, 100, 205,000, 20,500, 500 and 100 data points per channel.

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.

Speed tests were performed where a comparison of network prediction and interpolation calculation time for 5,000 examples were measured. The tests were executed on 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 or interpolated EEG data. Their task was to determine whether each presented example

was real, or computer generated (fake). 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 as either being represented by 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 marked as 'Real' or 'Fake'. 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 minutes to complete the test part. The data was displayed in average montage, the amplitude could be adjusted 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 with different MAE are given in Fig 1. Each channel 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.

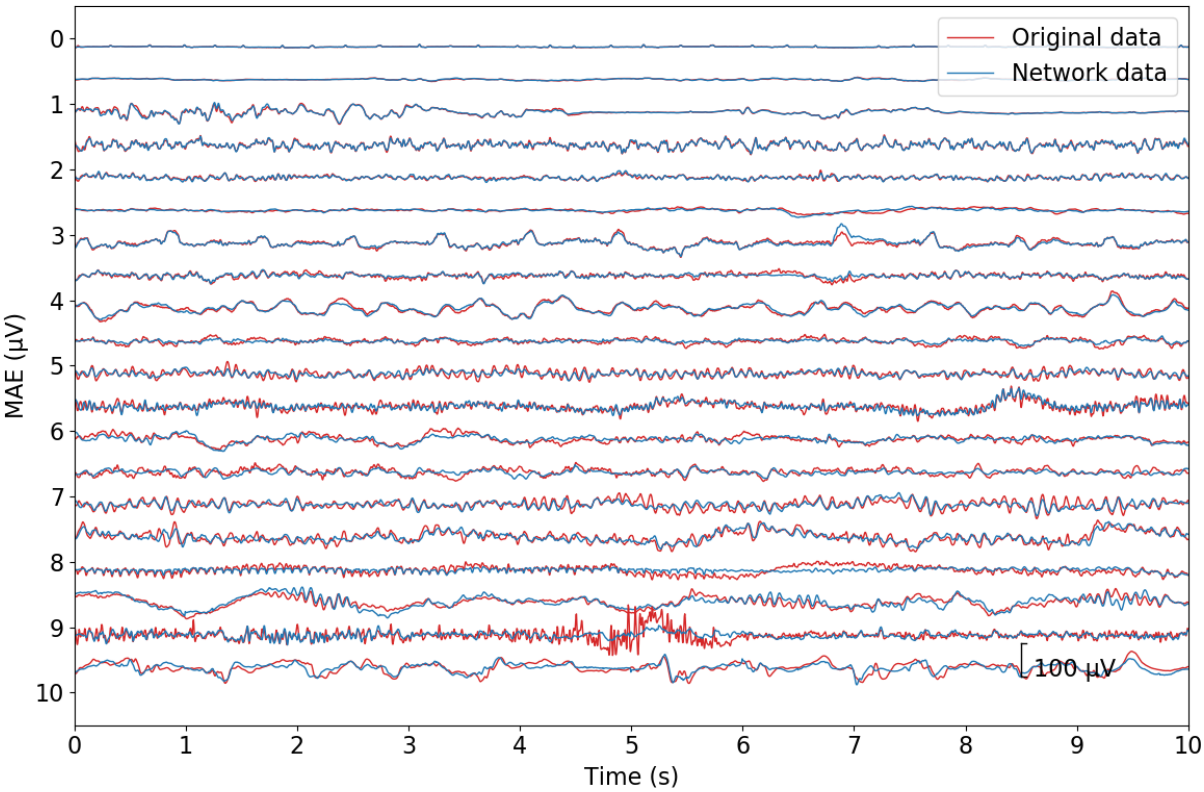

Fig. 1. Examples of recreated EEG signals with different MAE. Channels were randomly selected with increasing MAE, spanning the interval 0 – 10 µV. The original signal is in red and the signal recreated by GN1 is in blue.

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

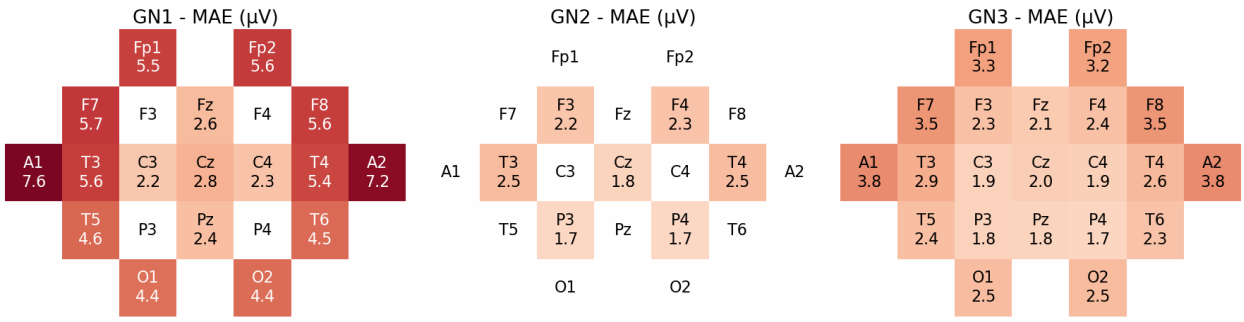

Fig. 2. Values of MAE for each electrode for the networks GN1, GN2 and GN3, left to right. The color code represents the MAE values, with white = 0 and dark red = 8 µV.

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 ill posed.

The results of the performance comparison are shown in Fig. 3. Overall, most measures showed a better performance for the convolutional networks compared to the spherical spline interpolation; 14 out of the 18 calculated values were significantly better. 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.

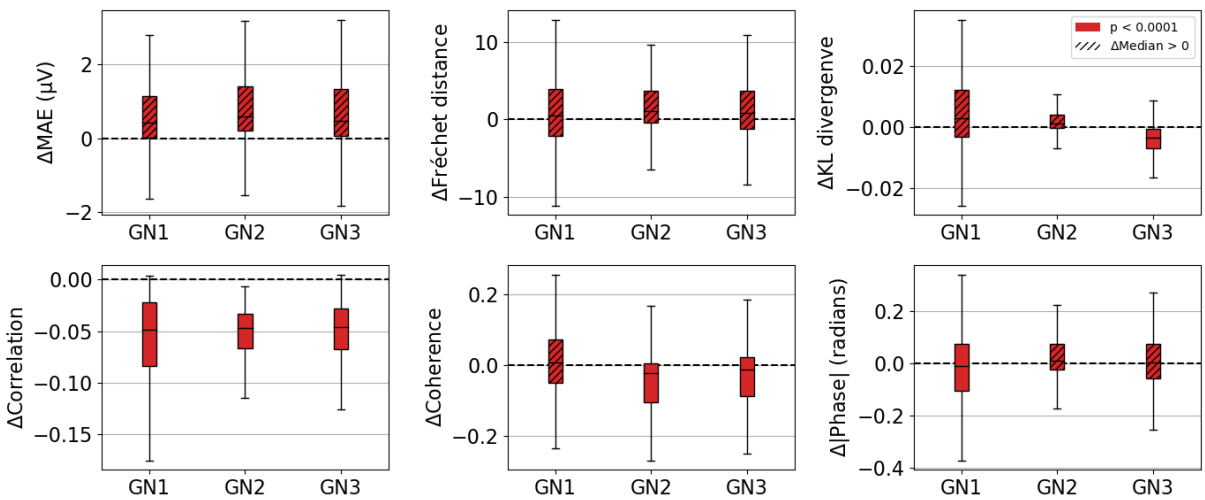

Fig. 3. The difference in the median between the convolutional networks and interpolation for all measures. The number of channels c per network were 17, 7 and 21 for GN1, GN2 and GN3, respectively. The number n of values per channel were 5,000, 500, 100, 100, 5,000×41 and 5,000×41 for MAE, FD, KL, R, C and P, respectively. Each box-and-whisker in the graphs were calculated from c×n values.

Performing predictions using a GPU gave processing times that were shorter than for interpolation (using CPU); Table 5. Predicting with CPU increased the processing times by a factor of 3 – 11, depending on network used.

Table 5

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.

	GN1	GN2	GN3
Network, GPU (s)	41	29	85
Network, CPU (s)	476	319	235
Interpolation (s)	113	41	95

3.3 Visual tests

The test took less than 10 minutes for all 5 raters. 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. 4), respectively. There was no difference between the real and network generated data, both performed better when compared to interpolated data.

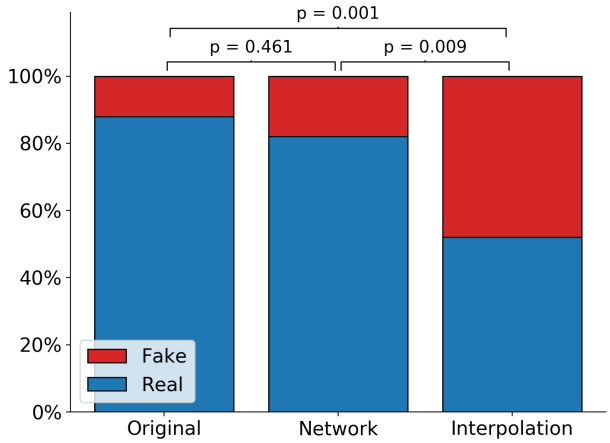

Fig. 4. Overall results of the visual assessment. The bars represent, from left to right: original data, data generated by GN1 and spline interpolation. Blue and red indicate the percentage of the examples that raters assessed as real or fake, respectively. Chi-square statistics were 10.8 (top bracket, p = 0.001), 0.5 (lower left bracket, p = 0.461) and 6.8 (lower right bracket, p = 0.009).

For the original data there were none, the GN1 network two and the interpolation eight examples where all raters were in full agreement of them being fake. These examples were visually inspected by one of us (MS) and the impression was that they looked cleaner than the originals, i.e., fewer artefacts, and less variation resulting in already monomorphic signals looking even more

monomorphic. This was in line with spontaneous comments by some raters after the test, and some also thought that some examples looked ‘too good’.

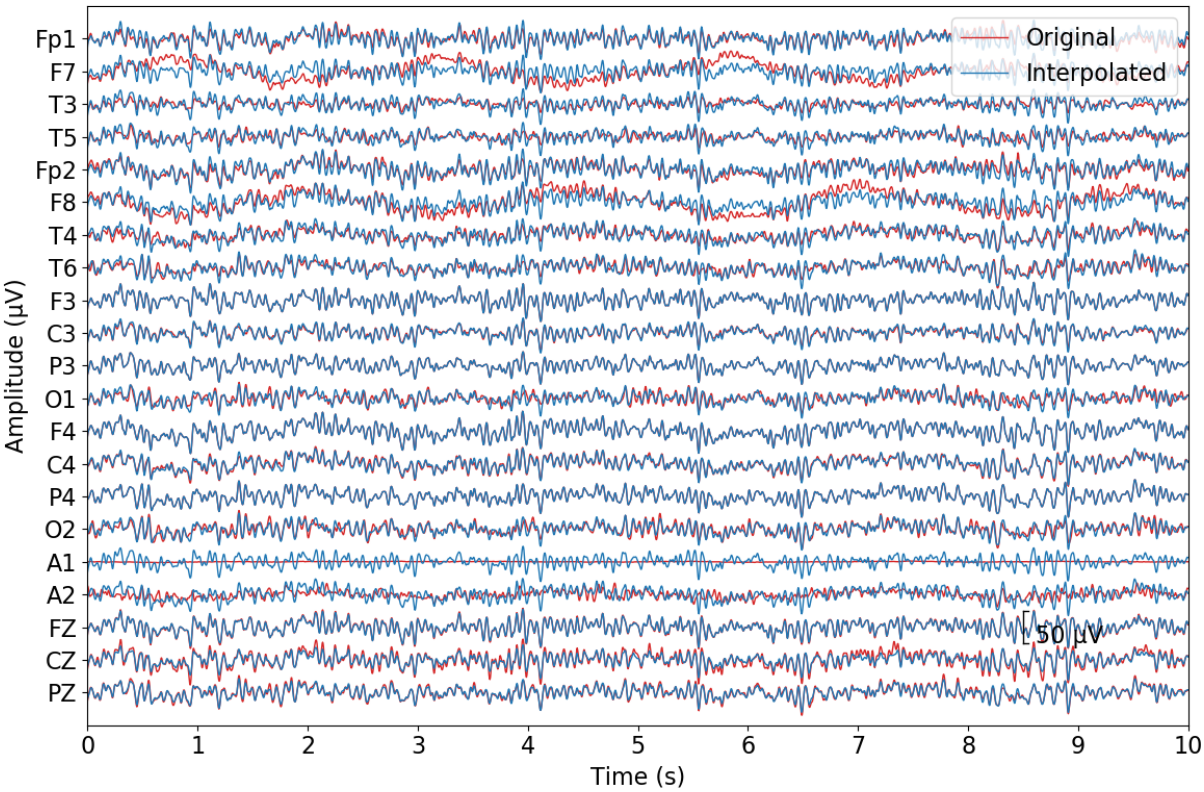

Fig. 5. Example of interpolated data (blue) and the original signals (red), where all raters voted ‘fake’.

3.4 The effect of the number of subjects

An increase in performance was seen by using more subjects for training. As can be seen in Fig. 6, the largest effect was seen for up to 100 subjects, after which a small additive effect was seen, however with the significance depending on which measure was being used.

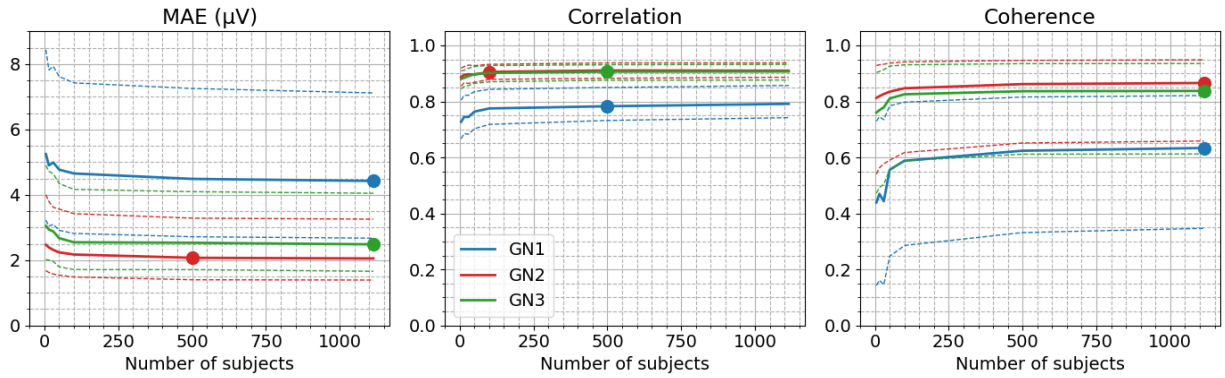

Fig. 6. MAE, correlation and coherence as a function of the number of subjects (5, 15, 30, 50, 100, 500, 1,114) used for training of the networks GN1 (blue), GN2 (red) and GN3 (green). Solid lines are the median, dashed lines the 25th and 75th percentiles, and the filled circles mark the points up to 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.

4 Discussion

We show that the convolutional neural networks can recreate EEG signals with a higher accuracy and produce signals with a more credible visual appearance than spherical spline interpolation. There are very few earlier studies using neural networks; the one we found used Wasserstein generative adversarial networks (Corley & Huang, 2018), which is hard to assess and compare given the large methodological discrepancies.

The amount of data used for our study was relatively large. However, the distributions of normal vs. pathological and of physiological variations in data character as well as occurrence of artefacts, were largely unknown and not controlled for during training and evaluation. Having control over the data distributions and using balanced data sets, may improve performance further. Though, assessing a data material of this size in such detail is an extensive undertaking.

Evaluating data of different character may reveal pros and cons of the method, (e.g., how different types of artefacts affect the results, and how wave transients and focal activity are reproduced). This kind of evaluation will be necessary if the method is to be used in clinical practice. As implied in the introduction, it has been speculated that predicting high frequency content and activity of more random character is more difficult compared to other kinds of activity. An example of a muscle artefacts, clearly not recreated by the network, is seen in Fig.1 at a MAE of 9 µV. Possibly in line with this, a posterior to anterior gradient is noted, with larger errors in the frontal quadrants (Fig. 2). Frontal and anterior temporal electrodes are likely to have more muscle artefacts on average, which may have contributed to this result. Eye movements, a low frequency, more organized artefact, also affects frontal electrodes the most and likely contributes to this. Also in the figure, an increasing error in radial direction is seen. This is probably an effect of increasing electrode distances, since it is reasonable to assume a larger error when predicting values over longer distances or from lower electrode density.

The networks were trained with an even distribution of examples regarding the number of subjects and the largest increase in performance was seen at 100 subjects (corresponding to 414 hours of data) with only slight increases thereafter. There was some variation regarding the significance depending on network and the measure used for evaluation. The significance 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 will also influence the results and how many subjects are needed to saturate the improvement.

In this work, the focus was not on testing a large variety of architectures and parameters. To train and test every possible network architecture and hyperparameter alternative is an intractable task – so there is no way of ever knowing for sure if there exist better choices than the best performing of the ones tested. Sometimes, small differences in settings may make all the difference but can

easily be missed. This means that there are probably better network architectures and hyperparameters so that the performance might be further improved.

Whether the processing times are acceptable will depend on the available hardware and the amount of data to be interpolated. Prediction using CPU for a standard recording of 20 min took 5.6 – 11.4 s with the computer used for the tests, which would be acceptable in practice.

Our demonstration shows that the networks can upsample and restore signals suggesting that the networks may be 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.

5 Conclusions

We demonstrate 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. 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, spatial asymmetries and how different artefacts affects the results. Processing times are acceptable for standard EEG exams. It may be worth-while using up to 100 subjects for training a network for this type of task. Further gains in

performance is possible by adding more subjects but may not be worth the effort if data is hard to acquire.

Disclosures and conflicts of interest

Declaration of interests: none.

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.

Acknowledgement

We would 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. et al., 2013. Automated EEG analysis of epilepsy: A review. *Knowledge-Based Systems,* Volume 45, pp. 147-165. https://doi.org/10.1016/j.knosys.2013.02.014.

Bach, S. H., He, B., Ratner, A. & Ré, C., 2017. *Learning the Structure of Generative Models without Labeled Data.* Sidney, Australia, Proceedings of the 34th International Conference on Machine Learning.

Corley, I. A. & Huang, Y., 2018. *Deep EEG super-resolution: Upsampling EEG spatial resolution with generative adversarial networks.* Las Vegas, Nevada, USA, IEEE EMBS International Conference on Biomedical & Health Informatics, pp. 100-103.

Eiter, T. & Mannila, H., 1994. *Computing discrete Fréchet distance,* Vienna, Austria.: Tech. Report CD-TR 94/64 Christian Doppler Laboratory for Expert Systems.

Felleman, D. & van Essen, D. C., 1991. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebal cortex,* Volume 1, pp. 1-47. DOI: 10.1093/cercor/1.1.1.

Fletcher, E. M., Kussmal, C. L. & Mangun, G. R., 1996. Estimation of interpolation errors in scalp topographic mapping. *Electroenceph clin Neurophysiol,* Volume 98, pp. 422-434. DOI: 10.1016/0013-4694(96)95135-4.

Freeden, W., 1984. Spherical spline interpolation-basic theory and computational aspects. *Journal of Computational and Applied Mathematics,* Volume 11, pp. 367-375. https://doi.org/10.1016/0377-0427(84)90011-6.

Gramfort, A. et al., 2013. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience,* Volume 7, pp. ISSN 1662-453X. DOI: 10.3389/fnins.2013.00267.

Hera, J. et al., 2017. Reduced electrode arrays for the automated detection of rhythmic and periodic patterns in the intensive care unit: Frequently tried, frequently failed?. *Clin Neurophysiol,* Volume 128, pp. 1524-1531. DOI: 10.1016/j.clinph.2017.04.012.

Hubel, D. H. & Wiesel, T. N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol,* Volume 160, pp. 106-154. DOI: 10.1113/jphysiol.1962.sp006837.

Jasper, H. H., 1958. The ten twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol,* Volume 10, pp. 367-380.

Kingma, D. P. & Ba, J., 2015. *Adam: A Method for Stochastic Optimization.* San Diego, International Conference on Learning Representations.

Kubota, Y., Nakamoto, H., Egawa, S. & Kawamata, T., 2018. Continuous EEG monitoring in ICU. *J. Intesive Care,* pp. 6:38. DOI: 10.1186/s40560-018-0310-z.

Kullback, S. & Liebler, R. A., 1951. On information and sufficiency. *Ann. Math. Statist.,* Volume 22, pp. 79-86. DOI: 10.1214/aoms/1177729694.

Lauritzen, S. L., 1974. Sufficiency, Prediction and Extreme Models. *Scand J Statist*, pp. 128-134.

LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep learning. *Nature,* Volume 521, pp. 436-444. DOI: 10.1038/nature14539.

LeCun, Y., Haffner, P., Bottou, L. & Bengio, Y., 1999. Object Recognition with Gradient-Based Learning. In: *Shape, Contour and Grouping in Computer Vision.* s.l.:Springer-Verlag, p. 319.

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. *IEEE International Joint Conference on Neural Networks,* 2004, Volume 4, pp. 2877-2882. DOI: 10.1109/IJCNN.2004.1381116.

Nahrstaedt, H. & C., L.-M., 2017. *holgern/pyedflib.* [Online] Available at: https://github.com/holgern/pyedflib [Accessed 21 09 2019].

Nunez, P. L. & Srinivasan, R., 2006. Human alpha rhythms. In: *Electric fields of the brain.* s.l.:Oxford University Press, pp. 12-19.

Nunez, P. L. & Srinivasan, R., 2006. Neocortical dynamics, EEG and cognition. In: *Electrical fields of the brain.* New York: Oxford University Press, pp. 510 (+369, 401-402).

Obeid, I. & Picone, J., 2016. The Temple University Hospital EEG data corpus. *Frontiers of Neuroscience,* Volume 10, pp. 1-5. DOI: 10.3389/fnins.2016.00196.

Perrin, F., Pernier, J., Bertrand, O. & Echallier, J. F., 1989. Spherical splines for scalp potential and current density mapping. *Electroenceph clin Neurophysiol,* Volume 72, pp. 184-187. https://doi.org/10.1016/0013-4694(89)90180-6.

Perrin, F. et al., 1987. Mapping scalp potentials by surface spline interpolation. *Electroenceph clin Neurophysiol,* Volume 66, pp. 75-81. https://doi.org/10.1016/0013-4694(87)90141-6.

Reaz, M. B. I., Hussain, M. S. & Mohd-Yasin, F., 2006. Techniques of EMG signal analysis: detection, processing, classification and applications. *Biol. Proced. Online,* Volume 8(1), pp. 11-35. DOI: 10.1251/bpo115.

Roy, Y. et al., 2019. Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.,* 16 051001(5), pp. https://doi.org/10.1088/1741-2552/ab260c.

Schomer, D. L. et al., 2018. Recording principles: Analog and digital principles; polarity and field determinations; multimodal monitoring; polygraphy (EOG, EMG, ECG, SAO2). In: *Niedermeyer's Electroencephalography.* s.l.:Oxford University Press, pp. 104-105.

Shapkin, A., 2012. *Reading and saving of data in the EDF+.* [Online] Available at: 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,* Volume 87, pp. 1985-195. https://doi.org/10.1016/0013-4694(93)90018-Q.

Tatum, W. O., Reinsberger, C. & Dworetzky, B. A., 2018. Artifacts of recording and common errors in interpretation. In: *Niedermeyer's Electroencephalography.* s.l.:Oxford University Press, pp. 267-279.

Völker, M. et al., 2018. *Deep transfer learning for error decoding from non-invasive EEG.* GangWon, South Korea, The 6th International Winter Conference on Brain-Computer Interface. DOI: 10.1109/IWW-BCI.2018.8311491.