Proceedings of the
9th Graz Brain-Computer Interface Conference 2024
10.3217/978-3-99161-014-4-069

THE GOOD, THE BAD, AND THE UGLY OF IEEG SIGNALS: IDENTIFYING ARTIFACTUAL CHANNELS USING CONVOLUTIONAL NEURAL NETWORKS

Z.V. Freudenburg[1], Y. Zhing[1], M. P. Branco[1], N. Ramsey[1]

[1] Brain Center, University Medical Center Utrecht, Utrecht, The Netherlands

E-mail: Z.V.Freudenburg@umcutrecht.nl

ABSTRACT: Intracranial electroencepghalography (iEEG) signals have established themselves as a key tool for studying human brain function due to its distinct combination of high spatial and temporal precision. The use of both cortical surface and stereo-EEG in effective epilepsy treatment has allowed researchers to study electrophysiology throughout the brain in relatively large numbers of subjects. This provides an opportunity to overcome, the sparse and varied nature of the brain tissue sampling inherent to the clinical use of iEEG by aggregating data across many subjects. Essential to the success of large-scale data aggregation is the efficient and robust identification of recording channels that are dominated by 'noise' or artifacts introduced by the recording environment or hardware failure. Here we test the effectiveness of training a convolutional neural network (CNN) for this purpose across multiple types of iEEG recordings. We conclude that a small CNN trained on hand labeled data from a small set of subjects can be applied to identify artifactual channels.

INTRODUCTION

Electroencephalography (EEG) allows for the recording of electrical signals generated by brain function and as such provides a precise measure of the temporal dynamics of brain function. However, extra-cranial EEG presents many challenges in terms of precisely locating independent neural sources of this activity. In the past decades the need to localize brain activity at the spatial resolution of tens of millimeters and with ms temporal precision to facilitate the localization seizure focus sites for medication resistant epilepsy treatment has led to the intracranial implantation of electrodes (iEEG) either on the cortical surface, often referred to as Electrocorticography (ECoG), or along shafts probing cortical and subcortical areas, often referred to as stereo-EEG (sEEG) [1]. The iEEG's use in epilepsy treatment and increased use in awake craniotomies for functional localization during brain tumor resection offers a unique opportunity to study the brain function of many humans performing a variety of motor and cognitive tasks.

However, due to the nature of the clinical setting in which iEEG is often recorded, the locations that are measured from only sparsely sample the brain and vary widely in number and location between subjects. Hence, showing reproducibility of results over humans on a whole brain scale for iEEG requires the aggregation of data across tenths or hundreds of subjects.

To facilitate this scale of iEEG data aggregation a robust and efficient method for identifying iEEG signals that are dominated by artifacts or noise introduced by hardware failure or fed by the environment is needed. Often such noise screening relies on the evaluation of experienced iEEG clinicians and researchers. This process is generally quite labor intensive, subjective, and not standard between centers or experts. Here we explore the effectiveness of training a deep learning model to do this.

Multiple groups have use the also attempted to use deep learning for noisy EEG channel selection. One approach is to apply thresholds to certain statistics computed from the signals. For example, APP [2] uses correlation and dispersion, FASTER [3] utilizes correlation, variance, and the Hurst exponent, Automagic [4] employs an in dependent component analysis-based artifact correction method, CTAP [5] calculates log relative variance and compares it to the median, and so on. Additionally, there are unsupervised methods. For instance, the Local Outlier Factor algorithm [6] identifies bad channels relative to the local neighboring channels, while the bad-by-RANSAC method [7] uses good channels to predict other channels and deems the channel poorly predicted by others as bad. Furthermore, supervised neural networks have also been utilized [8].

Yet, the transference to iEEG of these methods seems to be limited, with fewer reports about bad channel detection. The common method is to calculate statistics over the signals and input these statistics into machine learning methods. For example, in [9], the ensemble bagging classifier was applied to sEEG data, achieving the best accuracy of 99.77% across 110 subjects. In [10], multiple machine learning methods were tested on pigeons' ECoG data, including the K-Nearest Neighbors Algorithm (kNN), Support Vector Machine (SVM), Random Forest (RF), and others, with the best F1-score of 0.9089 achieved using RF and Synthetic Minority Oversampling Technique (SMOTE) to address the imbalanced dataset.

We chose for a Convolutional Neural Network (CNN) architecture because of the its proven ability learn EEG and iEEG signal filters at the lower level of more complex deep networks such as HTNet [11]. This allows us not to relay on predefined derived signal features

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-069

while keeping the network relatively simple, since we are not interested in differences in electrophysiological patterns or there spatial distribution on the cortex, but single channel level identification of non-electrophysiological (noisy or artifactual) signal. We apply this model to the three different iEEG recording modalities discussed above (clinical-ECoG, high-density (HD)-ECoG, HD, and sEEG) from a large number of individuals.

MATERIALS AND METHODS

Data: Data from 47 patients implanted with iEEG electrodes for the purpose of drug resistant epilepsy treatment at the University Medical Center in Utrecht were used is this study. The study was approved by the Medical Ethical Committee of the University Medical Center Utrecht in accordance with the Declaration of Helsinki (2013). The patients had either sEEG or clinical scale ECoG electrodes implanted according to clinical needs and gave written informed consent to participate in research tasks and will be referred to as subjects in this work. A subset of the subjects gave additional consent to have HD-ECoG implanted solely for research purposes. Data from and additional 3 patients undergoing an awake craniotomy for tumor removal, in which HD-ECoG grids are briefly placed on the exposed cortical surface, who also consented to performing brief research tasks were also included.

Data from clinical-ECoG implants were recorded from implanted grids and strips of evenly spaced platinum electrodes with an inter-electrode distance of 10 mm and a 2.3mm exposed recoding surface. Implanted sEEG props had 8-15 platinum-iridium cylinder contact points of 0.8 mm diameter and 2mm height with a 1.5 mm inter-contact distance. The HD-ECoG grids used were equally spaced grids of 32-128 platinum with 1.3 mm exposed surface diameter and an inter-electrode distance of 3 or 4 mm.

A total 96 data sets from 50 subjects performing one of 19 different cognitive tasks. Tasks range from simple relaxation without movement to overt and imagined movements to overt and covert speech. Subject data were organized into groups and based on the type of iEEG implant used to facilitate exploration of implant type on noisy signal detection (see Table 1). Furthermore the Clinical-ECoG group was split into an adult group (Ca) and a child group (Cc) test for an age effect on noise detection and the HD-ECoG group (HD) was divided by the recoding setting since a subset of this data was recorded in the operating room (OR) and not in outside the Intensive Epilepsy Monitoring Unit (IEMU) like the remaining subjects because it is known the OR has more noise sources and recordings are made while the electrode grids are still exposed to the air. In the case of the IEMU recordings are made after the scull has been replaced. The sEEG group (sE) was not further subdivided.

Table 1: Data Groups

Grid type	Sub-group	Subject count	Channel count*	Ratio good:bad
Clinocal-ECoG (Ca + Cc)	adult	11	1500	44:1
	child	13	1900	104:1
sEEG (sE)	adult	10		
	child	4 (14)	2700	43:1
HD-ECoG (HDe + HDor)	IEMU	9		
	OR	3 (12)	2000	33:1

(* = rounded to 100s)

Preprocessing and labeling: The recordings are from over a span of 20 years and have different frequencies, ranging from 512Hz to 2048Hz. To ensure that the bad channels exhibit similar patterns, all the recordings are down sampled to the lowest frequency, 512Hz.

For every subject, 10 minutes of recordings are included. For some subjects 5 minute recordings from the beginning are taken from two task files.

In this study we considered as noisy (bad) channels those that are clearly distinct from others in terms of signal content and that would likely distort the signals of non-noisy (good) channels when included in common average re-referencing (CAR). Bad channels were identified by visual inspection by two independent people (author 2 and author 3). For that we visually inspected both the raw-voltage signals and the power-spectrum (1/f, after removal of line-noise and it's harmonics) of every channel in one data file. Channels that had a deviant voltage amplitude compared with other channels in the same file (average voltage amplitude lays between -500 and 500 mV), excessive amount of line-noise, or recurrent large voltage fluctuations throughout the 5 minutes of data, were labelled as ‘bad’. Besides determining bad channels, we also identified borderline-bad channels, that would not be considered as 'bad channel' by an expert but could potentially be labelled as such by the algorithm. These channels were labelled as 'maybe'. The remain channels were labeled as ‘good’.

In this work a binary classification model is used, since we want the model to learn patterns of good and bad channels and be able to classify uncertain channels afterwards. Therefore, the ‘maybe’ channels are not included in the training or testing set.

The ratio of good vs bad channels in the data is very high, ranging from 44:1 to 104:1 (see Table 1). During the training, such severely imbalanced data harms the performance of the model. To reduce such effects, a down-sampling method was used in the training set to reduce the ratio of good vs bad channels to 2. The average number of good channels needed per subject need to achieve the 2:1 with the number of bad channels in the training set is calculated. Then random sampling is done to reach the designated amount of good channels per subject. By down-sampling, the amount of data in the training set also decreases, reducing the training time drastically.

According to clinical expertise, a window of 30 seconds in a channel contains enough information to show baseline patterns and classify channels as good or bad. In

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-069

addition, the level of noise can change over time due to head movements or medical operations. Splitting a channel into windows allows the model to classify windows in the same channel differently to account for possible changes in characteristics over time. Thus, each recording is chopped into around 10 30-second windows. These windows do not overlap to prevent data leakage from the training set into the testing set. For training all windows from a channel are given the label of that channel.

Model and Training: The model is a shallow Convolutional Neural Network (CNN) consisting of 2 convolutional layers with 55 weights followed by 2 linear layers with around 500k weights. To prevent overfitting, Batch-normalization, MaxPool, and Dropout layers with a rate of 0.2 are included. The input of the model is the 30-sec window, i.e., 30 seconds * 512Hz = 15360 nodes. The structure of the model was derived from experimentation with a training set containing one or two subjects from the Ca group and applied for the rest of the analysis.

Performance Evaluation: The model produces a prediction for each window of data. However, the prediction performance of the model for each window is not evaluated window-wise. Instead, the prediction of a channel is calculated by thresholding on the percentage of its windows that are predicted to be bad.

We chose the Matthews Correlation Coefficient (MCC) as our performance evaluation metric because it has been shown to be more robust and reliable among common metrics for imbalanced datasets [12]. MCC is a metric that summarizes a confusion matrix and computes the correlation between ground truth and predictions. The MCC is bounded from -1 to 1, with -1 indicating all predictions are wrong, 0 implying nearly random predictions from the model, and 1 indicating all predictions are correct.

Since the channel level performance of the model will depend on the chosen windowing threshold discussed above a metric for choosing this threshold is needed. In this context this comes down to choosing the best dividing line between the distribution of percent of windows labeled as bad for the set of bad channels labeled by the experts as bad and those labeled as good. For this we use Brent's method [13] because it works better than other methods for finding the border between two distributions when those distributions are multi-model. Brent's method is a numeric method to find the local minimizer of a certain function. Here, the parameter to minimize is the threshold, bounded by 0 and 1. The function calculates the negated MCC given the threshold and labels of the channels. Brent's method returns the threshold that maximizes MCC.

In this work we performed both within group and between group training-testing comparisons. To evaluate within group performance a leave-one subject-out (LoO) training testing approach was used. For this all channels from a single subject are left out of the training set and included in the test set and this is repeated to gain prediction results for every subject in the testing group. This means that when we reports results for training and testing on different sub-groups, if the training group sub-groups overlap with subgroups in the testing group the LoO method was used for subjects within the overlapping sub-groups.

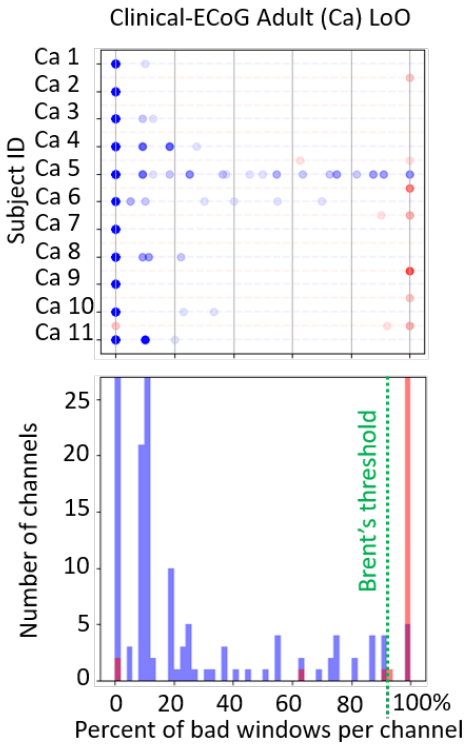

Figure 1: Leave-one subject-out (LoO) results for the Clinical-ECoG Adult (Ca) group. Top: Plot of % of windows classified as noisy (x-axis) for all channels of each subject (y-axis). Red and blue bad and good labeled channels respectively. Bottom: Histogram of number of channels (y-axis) with a certain percentage of windows predicted as bad (x-axis). The red and blue bars represent channels labeled as bad good respectively. The green vertical dotted line indicates the threshold found with Brent's method.

RESULTS

The distinction between gad and bad varies over iEEG groups As Figures 1 and 2 illustrate the distinction between the good and bad channels in terms of the percentage of windows classified as bad varies considerable between iEEG date groups. When channels are clearly bad or good most of the time and the model is able to learn a clear distinction between the two signal types you would expect to see the distributions found for the Ca group in Figure 1. Here we see that almost 100% of the windows are classified as bad for most of the channels labeled as bad and often not more than 30 % of the windows from good channels are classified as bad. This means that a wide range of thresholds (30%-90%) will give similarly good MMC scores. In this case a very

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-069

conservative threshold of 92% that allows for good channels to have a lot of bad windows is found by Brent's method to give the optimal MMC score of 0.84 (see Table 2, first row). This effectively means that only 5 good channels 4 channels labeled as noisy (out of ~1500) are misclassified. However, as Figure 2 shows, in the case of the sE group both the distributions of good and bad labels are broader. Meaning that either the network has a harder time distinguishing good windows from bad windows or that the amount of noise in the sEEG signals fluctuates over time causing good channels to have noise at times and bad channels to be less noisy at times. In this situation the optimal threshold really need to balance the false positives (FPs) and false negatives (FNs) to reach the optimal MMC, which was found to be 0.32 at 68% (Table 2) in this case. It should be noted that this MMC is still well above 0 and the good and bad distributions for sE are still distinct from each other allowing for thresholds to be set that can reduce either the number of FPs or FN to almost 0, but not both.

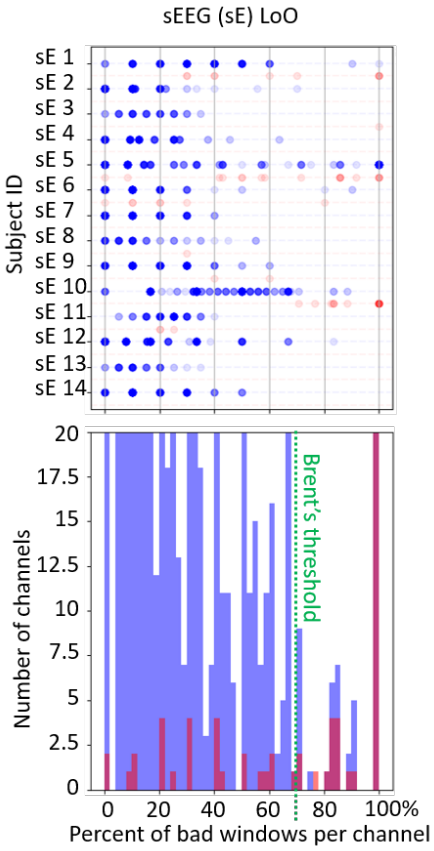

Figure 2: Leave-one subject-out (LoO) results for the Clinical-ECoG sEEG group. Top: Plot of % of windows classified as noisy.

In general bad electrode classification works well for Clinical-ECoG As can be seen in Table 2 all analysis that involved training and testing on Clinical-ECoG groups performed well. The Cc within group test showed and even higher MMC score (0.94) that that of Ca and even when training on Ca and testing on Cc an MMC score of 0.89 was achieved. This indicates that there is little difference between the good and bad signal characteristics between channels implanted in adults and children. This is further supported by the fact that training on Ca and Cc also gave high MMCs of 0.92 and 0.82 when testing on Ca and Cc respectively.

Table 2: MMC and Brent threshold across different training and test groups combinations.

Test Group	Train Group	MCC	Brent's threshold
Ca	Ca	0.84	92%
Cc	Cc	0.94	85%
sE	sE	0.32	68%
HD	HDe+HDor	0.10	48%
HDe	*HDe+HDor*	*0.43*	*94%*
Cc	Ca	0.89	85%
Ca	Ca + Cc	0.92	70%
Cc	Ca + Cc	0.82	70%
sE	Ca + Cc	0.66	76%
HD	Ca + Cc	0.23	77%
sE	Ca+Cc+sE	0.41	85%

Distinguishing the bad from the good in sEEG is harder but promising While the sE group performance discussed above is lower than that of the Clinical_ECoG groups introducing data from these groups into the model training does improve its performance on the sE group to an MMC of 0.41 (bottom row, Table 1). In fact, only training on the Ca and Cc groups improve the performance on the sE group even further to 0.66. This is a marked improvement and inspection of the percentage bad windows distributions (Figure 3) indicates that by training on channels with clearer bad vs. good signal distinctions the model was able to more clearly separate good and bad sE channel windows. While there are now many bad channels with none of their windows classified as bad, the number of bad channels with >80% of windows classified as bad was relatively unaffected and the number of good channels with >50% of their windows classified as bad decreased greatly.

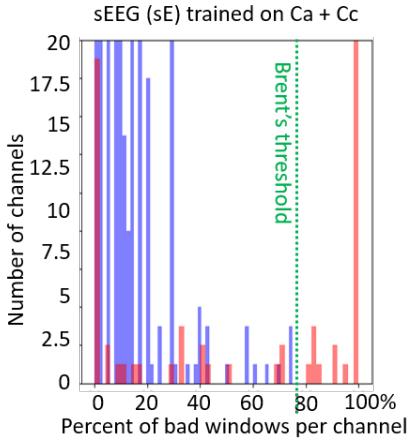

Figure 3: sEEG window classification based on training

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-069

on Clinical-ECoG.

HD-ECoG presents a challenge to automatic noisy channel detection While still above 0 the MMC for the HD group was only 0.1 (Table 2) and the fact that the optimal threshold was around 50% (48%) and the top plot in Figure 4 indicate that the good and bad distributions were very mixed. In fact, most of the good and the bad labeled channels had around 50% of there windows classified as bad. This could indicate that the amount of noise in the HD-ECoG channels changes a lot over time. It is worth noting that this is especially the case for the sub-group of subjects recorded in the operating room setting (HDor). When excluding the HDor subjects for testing, the performance increases to and MMC of 0.43. While training on the Clinical-ECoG groups does improve performance on the HD group as a whole similar to the sE group, this improvement is mostly due to the better distinction on the HDe sub-group (Figure 4, bottom plot). In the case of HDor subjects, 0% of almost all channels are classified as bad. This is surprising since the HD group has the lowest good to bad channel ratio (Table 1) and the HDor sub-group has a the majority of the bad channels in this group, as can been seen by the large number of red circles for the HDor subjects in the top plot of Figure 4.

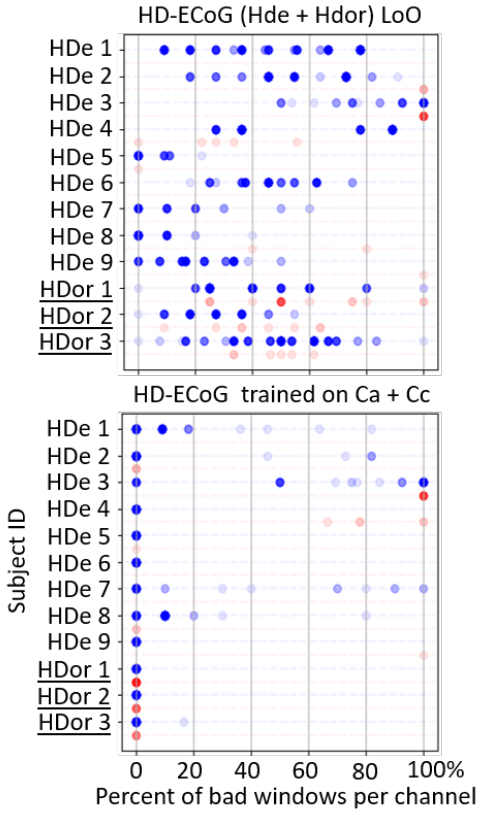

Figure 4: HD-ECoG noisy window classification for within group vs. Clinical-ECoG group training. Top: Plot of % of windows classified as noisy. Bottom: Plot of % of windows classified for the HDe and HDor groups after training on the Ca + Cc groups.

DISCUSSION

In general we found that our simple model worked well for identifying noisy artifactual signal in clinical-ECoG data and that there is no need to treat signals from children as different from those of adults.

In addition, while we found the our approach was less effective regaurding sEEG signals, we did see that performance on sEEG channels can be improved by applying a model trained on clinical-ECoG channels, where there is a larger separation between channels with artifactual signals and those with clean electrophysiological signals.

The results suggest that the difficulty in identifying noisy sEEG channels could be due to a larger variance in electrophysiological signal properties found in sEEG signals. When training on sE the model showed a broad distribution for both bad and good channels. It is known that the electrophysiology of sub-cortical regions differs from that of the cortical surface and the model has no knowledge of this while the expert labeler does. One trend is that the amplitude in signal generally decreases with the depth of the implanted electrode. The expert may consider this when judging a channel to be good even if it has low amplitude. This would indeed make it harder to train on sEEG channels as there will be a larger variance in the types of patterns associated with channels labeled as good. However, the fact the model trained on clinical-ECoG greatly increase the number of sEEG channels with < 20% of windows labeld as bad without decreasing the number of labeled bad or increasing the number of labeled good channels with > 80% of widows predicted as bad shows that relative to clinical-ECoG most sEEG channels contain similar data. This is promising for goal of agrigating sEEG data into clinical-ECoG data sets.

HD-ECoG presented the hardest challenge to accurately predicting bad channels. This was especially true for the sub-group of subjects recorded in the operating room. As opposed to sEEG this is likely due to an increase in the variety of types of artifactual signals in these recording compared to clinical-ECoG. This group had the lowest good to bad channel ratio and hence the largest number of labeled bad channels. The fact that the HD-electrodes are smaller means they will have larger impedences and potentially pick up more external noise though the wire connecting them to the amplifiers. In addition it is known that the recordings from the operating room in the HDor group will be more sensitive to external noise because during the recording the scull is open, unlick in the IEMU where the skull has be replaced and the wound sealed, before recording. This provides more direct exposure to external electrical sources. In addition the fact that most labeled bad and good channels have around 50% of their windows predicted as bad suggests that there may be artifacts that are transient in nature and not consistent throughout the recording. Taken together this suggests that predicting artifactual signal in HD-ECoG will likely require larger set of correctly labeled data to train on. It should be noted that often HD-ECoG implants are aim

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-069

towards BCI use and thus longer term implantation where additional work in identifying and labeling artifactual signal is justified.

Limitations The main limitation of this work is that ground truth labeling was based on subjective evaluation of channel as a whole. This is sufficient for large scale studies, but not a true separation of electrophysiological signals from external noise. Each channel is not purely noise or brain signal at any one time point. Thus, it could also be beneficial to allow for prediction of a percentage or probability of noisiness as the recorded signal is almost always a combination of both.

However, this proportion is hard to compute and label. This is a limitation inherent to all work on distinguishing artifactual signal from electrophysiological signal in that work on understanding the true ground truth of what parts of recorded iEEG signals are pure reflections of electrophysiology is still very much ongoing.

One approach to overcome this would be to train with as good as possible pure electrophysiological signals and add known amounts of simulated noise. In this way there would at least be ground truth for the know added signal artifacts and knowledge as bout simulating artifacts caused by known noise sources is much better due to the vastness of the field of electronics.

Furthermore, the models could be improved by allowing for additional types of labels for channels such as the cortical or subcortical region where the electrodes are located and/or the type of cognitive task the subject is performing as these factors are known to influence iEEG signal features. In this way, a model trained to identify noisy signal characteristics could also be used to specify what signal features constitute 'normal' iEEG signal from different parts of the brain. This suggests another possible avenue for future work, which would be to explore the signal features encoded in the deeper layers of such a CNN.

CONCLUSION

In conclusion we feel this work is encouraging for studies aimed at large scale data aggregation over many subjects and multiple institutions in that is shows the feasibility of automating the identification of channels that can be safely included in the analysis and which ones should be excluded.

REFERENCES

[1] Herff C, Krusienski DJ, Kubben P. The Potential of Stereotactic-EEG for Brain-Computer Interfaces: Current Progress and Future Directions. Front Neurosci. 2020;14:123

[2] da Cruz JR, Chicherov V, Herzog M, Figueiredo P. An automatic pre-processing pipeline for EEG analysis (APP) based on robust statistics. Clin. Neurophysiol. 2018; 129:1427-1437

[3] Nolan H, Whelan R, Reilly RB. FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. J Neurosci Methods. 2010;192:152-62.

[4] Pedroni A, Bahreini A, Langer N. Automagic: Standardized preprocessing of big EEG data. Neuroimage. 2019;200:460-473

[5] Cowley BU, Korpela J. Computational Testing for Automated Preprocessing 2: Practical Demonstration of a System for Scientific Data-Processing Workflow Management for High-Volume EEG. Front Neurosci. 2018;12:236

[6] Kumaravel VP, Farella E, Parise E, Buiatti M. NEAR: an artifact removal pipeline for human newborn EEG data. Dev. Cogn. Neurosci. 2022; 54

[7] Bigdely-Shamlo N, Mullen T, Kothe C, Su KM, Robbins KA. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. Front Neuroinform. 2015;9:16

[8] Kumaravel, V.P., Kartsch, V., Benatti, S., Vallortigara, G., Farella, E., Buiatti, M. Efficient artifact removal from low-density wearable EEG using artifacts subspace reconstruction, in: Proc. of 43rd Ann. Int. Con. of the IEEE Engineering in Medicine and Biology Society, 2021, 333–336

[9] Tuyisenge V, Trebaul L, Bhattacharjee M, Chanteloup-Forêt B, Saubat-Guigui C, Mîndruţă I, Rheims S, Maillard L, Kahane P, Taussig D, David O. Automatic bad channel detection in intracranial electroencephalographic recordings using ensemble machine learning. Clin Neurophys. 2018;3:548-554.

[10] Li M, Liang Y, Yang L, Wang H, Yang Z, Zhao K, Shang Z, Wan H. Automatic bad channel detection in implantable brain-computer interfaces using multimodal features based on local field potentials and spike signals. Comput Biol Med. 2020;116:103572

[11] Peterson S, Steine-Hanson Z, Davis N, Rao R, Brunton B. Generalized neural decoders for transfer learning across participants and recording modalities. J. Neural Eng. 2021;18:026014

[12] Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. 2021;14:13.

[13] Brent, P Chapter 4: An Algorithm with Guaranteed Convergence for Finding a Zero of a Function", Algorithms for Minimization without Derivatives, Englewood Cliffs, NJ: Prentice-Hall, 1973.

Published by
Verlag der Technischen Universität Graz