Deep Neural Networks for Automatic Classification of Anesthetic-Induced Unconsciousness

Konstantinos Patlatzoglou[1](✉), Srivas Chennu[1,2], Mélanie Boly[3], Quentin Noirhomme[4], Vincent Bonhomme[5,6], Jean-Francois Brichant[7], Olivia Gosseries[8], and Steven Laureys[8]

[1] University of Kent Chatham Maritime, Kent, UK
{kp356, sc785}@kent.ac.uk
[2] University of Cambridge, Cambridge, UK
[3] Department of Neurology and Department of Psychiatry, University of Wisconsin, Madison, WI, USA
[4] Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands
[5] GIGA - Consciousness, Anesthesia and Intensive Care Medicine Laboratory, University and CHU University Hospital of Liege, Liege, Belgium
[6] Department of Anesthesia and Intensive Care Medicine, CHU University Hospital of Liege, Liege, Belgium
[7] Department of Anesthesia, University of Liege, Liege, Belgium
[8] Coma Science Group, GIGA Consciousness, University and University Hospital of Liège, Liège, Belgium

Abstract. Despite the common use of anesthetics to modulate consciousness in the clinic, brain-based monitoring of consciousness is uncommon. We combined electroencephalographic measurement of brain activity with deep neural networks to automatically discriminate anesthetic states induced by propofol. Our results with leave-one-participant-out-cross-validation show that convolutional neural networks significantly outperform multilayer perceptrons in discrimination accuracy when working with raw time series. Perceptrons achieved comparable accuracy when provided with power spectral densities. These findings highlight the potential of deep convolutional networks for completely automatic extraction of useful spatio-temporo-spectral features from human EEG.

Keywords: Consciousness · Anesthesia · EEG · Deep learning

1 Introduction

In the United States alone, 60,000 people receive general anesthesia (GA) every day for surgery [1]. Despite the obvious fact that GA fundamentally modulates brain activity, brain monitoring is not routine practice in the operating room, and is limited to proprietary systems which have produced mixed results, in part due to considerable inter-individual variability [2]. Recent research into electroencephalographic (EEG) signatures of propofol-induced unconsciousness have highlighted the potential for improved brain monitoring [1, 3].

S. Wang et al. (Eds.): BI 2018, LNAI 11309, pp. 216–225, 2018.
https://doi.org/10.1007/978-3-030-05587-5_21

One of the challenges encountered in deploying novel EEG metrics of consciousness at the bedside is automation, in that they require expert analysis or interpretation of the data. To work towards addressing this challenge, we apply recent developments in artificial intelligence research, deep neural networks in particular, to the challenge of fully automated feature learning from EEG to detect states of unconsciousness due to propofol anesthesia. As there is no state-of-the-art deep learning model or reference dataset for EEG classification, we compare the performance of two widely used models, multilayer perceptrons (MLP) and convolutional neural networks (cNN), in their ability to discriminate states of unconsciousness from only 1 s of raw EEG data. With leave-one-participant-out-cross-validation, we show that cNNs achieve nearly 90% accuracy and significantly outperform MLPs, and generalize to data from participants unseen during network training.

2 Methods

2.1 Dataset Collection

The data used in this work were acquired from a propofol anesthesia study [4], in which the experimental design is described in detail. Briefly, the study was approved by the Ethics Committee of the Faculty of Medicine of the University of Liege, with participants giving written informed consent. Moreover, physical examination and medical history were obtained, in order to assure of any potential issues during anesthesia (e.g. pregnancy, trauma, surgery, mental illness, drug addiction, asthma, motion sickness).

Fifteen-minute spontaneous high-density electroencephalography (hd-EEG, 256 channel Hydrocel GSN) was recorded from 9 participants (mean age 22 ± 2 y, 4 males) during propofol anesthesia, at three different levels of consciousness, from fully awake, to mild sedation (slow response to command) and clinical unconsciousness (no response), as depicted in Fig. 1. Sedation procedure was monitored, while computer-controlled intravenous infusion was used to estimate effect-site concentrations of propofol. The level of behavioral consciousness was confirmed with the Ramsay scale, see [4] for details.

2.2 EEG Pre-processing

Minimal pre-processing steps were applied to the original data, in order to simulate a real-world scenario where deep learning could be applied to EEG data in real-time. Although raw EEG recordings tend to be noisy, the selection of the workflow was based on the notion of an automated feature extraction done by deep learning, along with a potential practical value of such implementation within a clinical context, where manual intervention and *a priori* knowledge of the signal would be infeasible.

Two different representations were extracted from the datasets, to compare the effects of using the raw time series versus a spectral representation. The latter has often been used in similar studies as a useful feature in EEG classification [5–8].

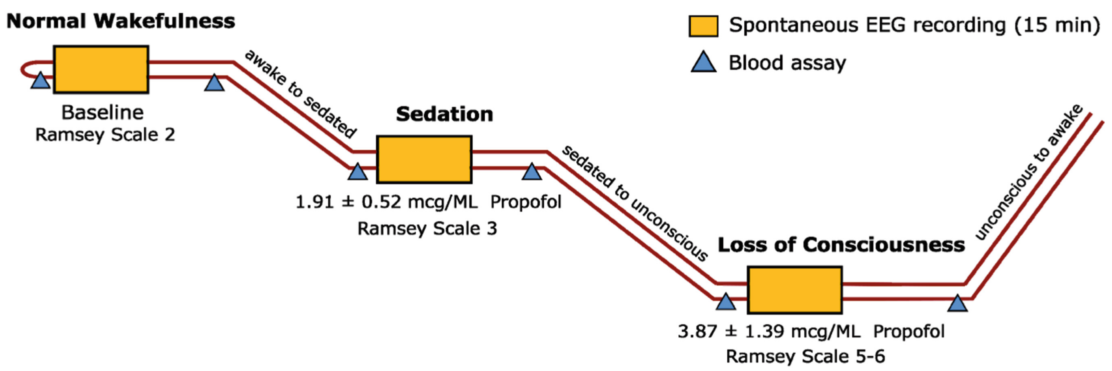

Fig. 1. Experimental design of the propofol anesthesia study. Participants underwent anesthetic induction into progressively deeper states of unconsciousness measured by behavior.

Raw Data Representation. For reducing the computational complexity of the deep learning pipeline, 20 electrodes of EEG data were examined, located as per the 10-20 system, namely: Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, Oz, and O2. Data were segmented into 1 s non-overlapping epochs and band-pass filtered between 0.5-40 Hz using a window FIR design (firwin, scipy). The vertex (Cz) electrode was the online reference, which was replaced by the average activity of all the 19 channels. Finally, the time series were down-sampled to 100 Hz, resulting in 100 samples per epoch. No manual artefact or bad channel rejection was performed other than the removal of the first 10 s of recording, which contained large unstable drifts. All pre-processing steps were implemented using the MNE-python library with default settings, unless specified otherwise.

Power Spectral Density Representation. To generate spectral representation of the EEG, raw data processed as above were submitted to the periodogram function (scipy) to obtain the power spectral density (PSD) of each channel and epoch. 201 points were used to compute the PSD, which resulted in 100 frequency bins (one-sided spectrum, dc coefficient removed). Importantly, this ensured that the dimensionality of the data was identical with both raw and PSD representations. The resulting dimension of each instance (epoch) was a 20×100 (channels x time samples/frequency bins) 2D-array for both representations.

Finally, the data were normalized by epoch using the scikit-learn library, before feeding them into the deep learning networks. This can be thought as normalizing the whole scalp activity for each epoch and participant independently. Although there are many ways to normalize the data (e.g. by time sample or by channel), this way was considered more appropriate in terms of its physical interpretation and practical application, as only data from the current epoch is required for applying the normalization.

2.3 Deep Learning Architectures

Two deep learning architectures were compared, as a way to investigate the suitability of such algorithms in classifying states of consciousness and extracting relevant features from the EEG. Convolutional neural networks (cNN) are a class of feed-forward networks that have become very interesting for end-to-end EEG research (both for analysis and interpretation of data) during the recent years. This architecture has shown to be very efficient in analyzing raw data (mostly from images), as it reveals spatial features across different levels of abstraction, using the convolution operation over local segments of the data [9]. In contrast, the Multilayer perceptron (MLP) network is a naïve implementation of a deep learning model, which can be used as a baseline for comparison (cNN can be thought as an MLP with a specialized structure).

Our aim here was not to optimize each network for the given task, but rather to compare them fairly, to reveal the computational advantages of each design. Hence the two models were compared with respect to their architectural sizes, which can be thought as the number of neurons/trainable parameters within each functional layer.

Convolutional Neural Network. The architecture of the cNN is a sequential model based on a simple design used in computer vision for hand-written digit classification (mnist example, Keras). The first functional layer (feature extraction) is a sequence of two convolutional layers, followed by a max-pooling and a dropout layer. The second functional layer (classification), consists of a fully connected layer, followed by a dropout layer and three softmax units (one for each conscious state). As a reference size, the original number of feature maps and hidden neurons were used, namely 32 for the 1st convolutional layer, 64 for the 2nd convolutional layer and 128 neurons for the 3rd dense layer. The patch window for max pooling was 2×2. Dropout rates were 0.25 and 0.5, respectively. Convolution windows were chosen with kernels 1×5 and 5×10 (1×1 strides), with the first layer only extracting temporal information (no padding used). Finally, all activation functions were relu units (except output layer). The model was trained using the categorical cross-entropy loss function and the Adadelta optimizer. Initialization of network weights was done with the Xavier uniform initializer. The cNN architecture is summarized in Fig. 2.

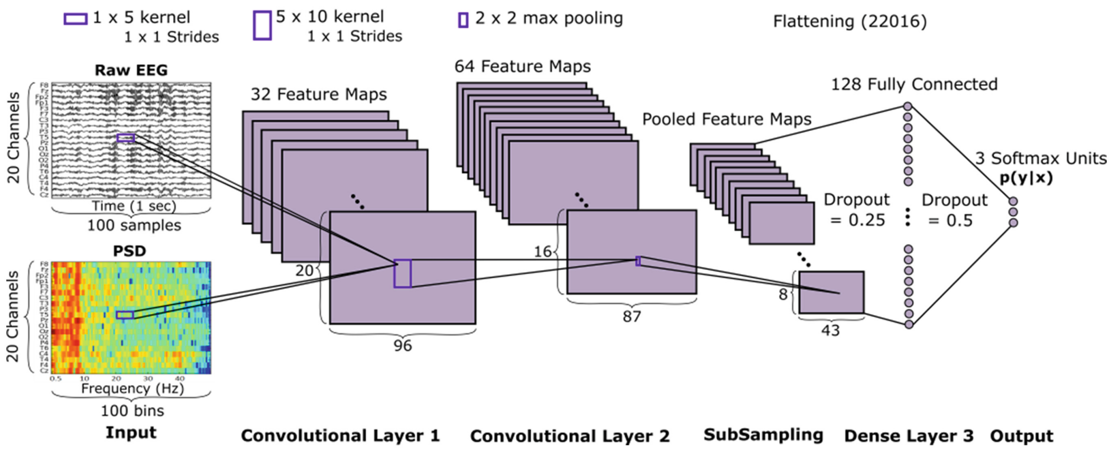

Fig. 2. Convolutional neural network architecture (reference size) for classifying the three conscious states: wakefulness, sedation, loss of consciousness. Raw EEG or PSD epochs were used as an input tensor.

Multilayer Perceptron. We employed a sequential MLP model designed to match the number of output neurons in each functional layer of the cNN (rather than equalising network layers). This ensured that the computational cost of each design was comparable in terms of training time. Both functional layers of the MLP consist of fully connected layers, followed by a dropout layer (2nd layer includes the three softmax units). The number of hidden units for the 1st layer was based on the number of neurons after the flattening in the cNN architecture (22016 for the reference size), while for the 2nd layer was kept the same. Activation functions, dropout rates and other model parameters during training were also kept the same with respect to the cNN. The MLP architecture is summarized in Fig. 3.

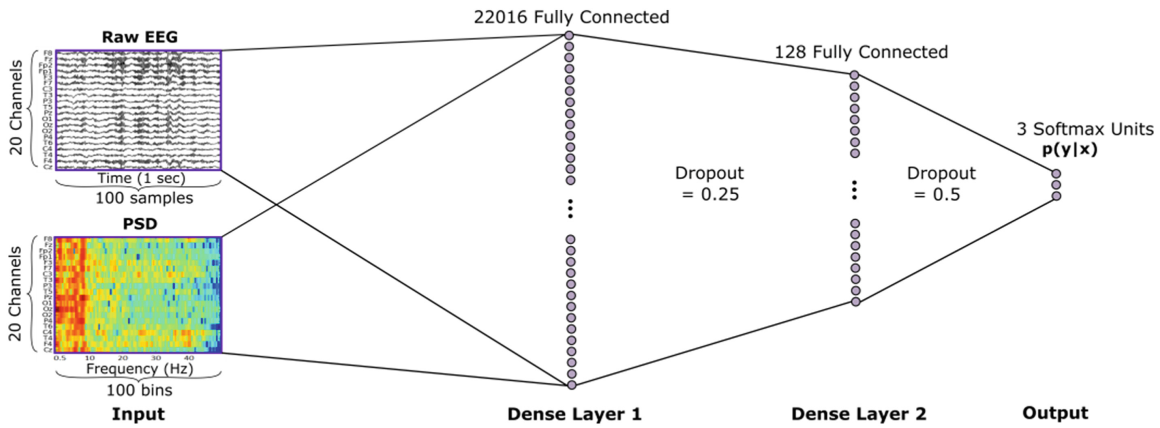

Fig. 3. Multilayer perceptron network architecture (reference size) for classifying the three conscious states. Raw EEG or PSD epochs were used as input tensors, after flattening the 2D-array into a 2000-dimensional vector.

2.4 Experiments

Twelve experiments were done in total for the 2 × 2 × 3 combinations of data representations (Raw vs PSD), deep learning architectures (MLP vs cNN) and 3 different network sizes – small, reference and large, in order to compare performance of the models. The number of feature maps and neurons of the fully connected layers for each architecture and network size are listed below (Table 1).

Table 1. Network Sizes

Network size	cNN	MLP
Small	(16, 32, 64)	(11008, 64)
Reference	(32, 64, 128)	(22016, 128)
Large	(64, 128, 256)	(44032, 256)

To evaluate model performance, EEG data were divided into training and test sets. Previous studies have divided data from each participant proportionally into training and test sets [6, 10, 11]. However, an ideal but hard goal would be to generalize and predict states of consciousness in unseen participants. With this goal in mind, leave-one-participant-out cross validation (LOPOCV) was used for the training and testing of the models, with each participant contributing 2700 instances on average (9 participants, 3 states, 15 × 60 1-s epochs ≈ 24300 total instances). Each instance was labeled with one-hot encoding as the target vector, indicating one of the three sedation states. Training was done with a batch size of 100 and for 10 runs (epochs). Models were evaluated by their accuracy, computed as the percentage of epochs correctly predicted in the left-out participant. All experiments were implemented in Python 3 using Keras/Tensorflow on a CUDA NVIDIA GPU (Tesla P100).

3 Results

3.1 Architecture Comparison

The results from our 2 × 2 experimental design (Raw/PSD *X* cNN/MLP) were similar for all three network sizes, and are summarized below. Reported figures and accuracies are for the reference size networks depicted in Figs. 2 and 3.

Raw Data. With raw EEG input, the MLP achieved an average accuracy of 75.45% across participants, with the cNN achieving 86.05% (Fig. 4). These accuracies are significantly higher than the chance level accuracy of 33.33%. Cross-entropy loss on the test set did not significantly decrease after the first epoch. Overall, the cNN was able to achieve better accuracies for each state of consciousness and participant.

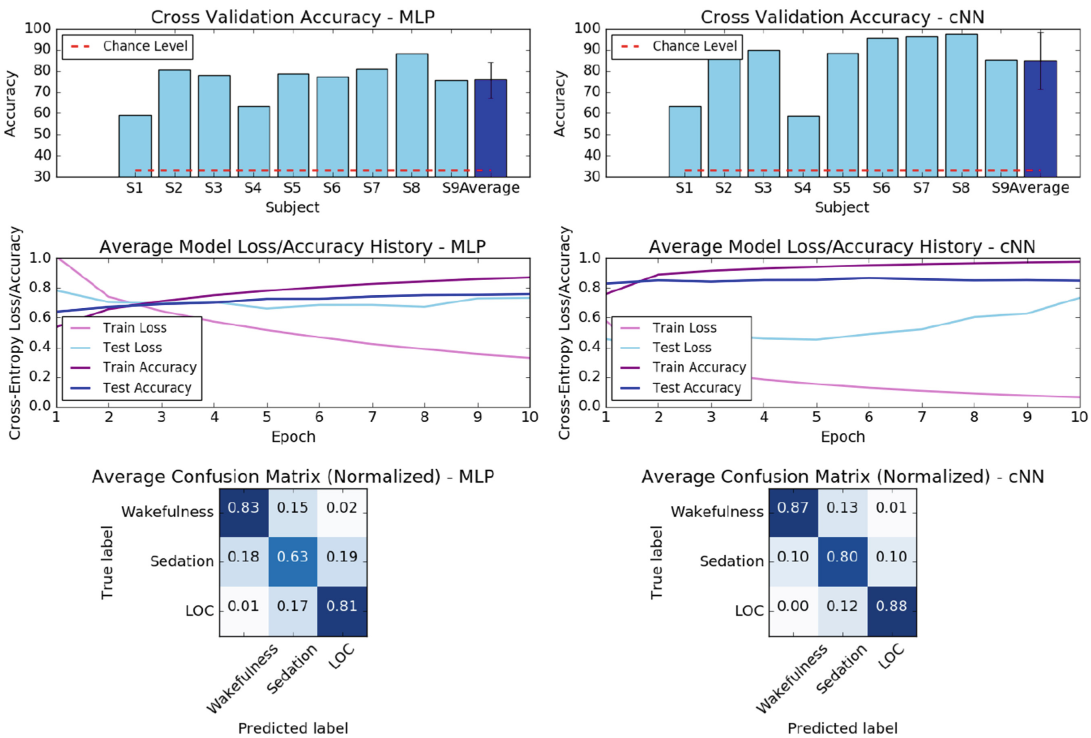

Fig. 4. MLP vs cNN (reference size) comparison for raw EEG classification of the three conscious states. Cross validation accuracies, average model loss and confusion matrices are shown for each architecture.

As seen in Fig. 4, the confusion matrices suggest that Wakefulness and LOC were not often confused. The intermediate state of Sedation was hardest to predict, due to individual variability in response to propofol. However, this would not present a problem in the clinical context, where anesthetic induction is much more rapid [10].

Power Spectral Density. With PSD input, the two architectures were equally capable in classifying states of consciousness (Fig. 5). In particular, the MLP performed better than when provided with raw time series as input, but the cNN did not (MLP: 83.4%, cNN: 87.35%). Importantly, cross-entropy loss revealed that the models converged faster using the PSD representation.

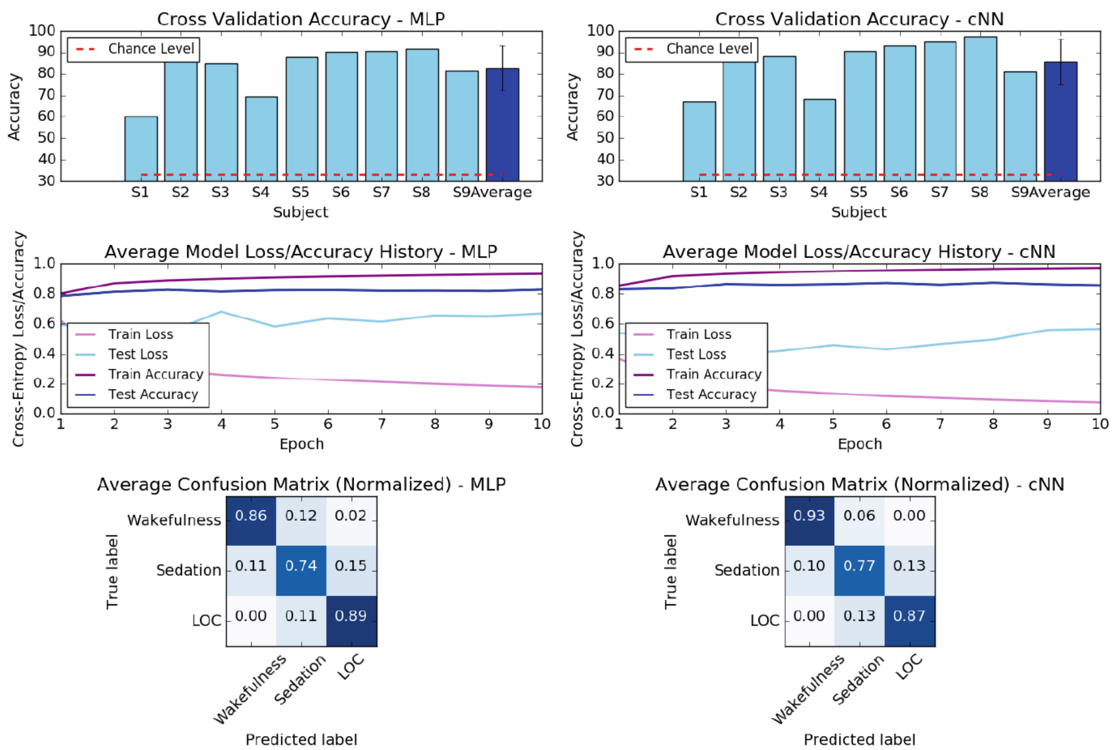

Fig. 5. MLP vs cNN (reference size) comparison for EEG classification of the three conscious states, using the PSD representation.

To understand the changes in the underlying EEG signal driving these accuracies, we visualized the PSDs in each state of consciousness (Fig. 6). As expected, we observed a decrease in alpha oscillations in Sedation, followed by the emergence of high-alpha oscillations during LOC.

3.2 Statistical Analysis – ANOVA Model

As a final step, a three-way ANOVA (type 2) was performed on the accuracies obtained in all twelve experiments across 2 architectures, 2 data representations and 3 model sizes, as summarized in Fig. 7 and detailed in Table 2.

The results of the ANOVA indicated that network architecture (cNN/MLP) was the strongest contributor to model performance ($F = 10.6$, $p = 0.0015$), while data

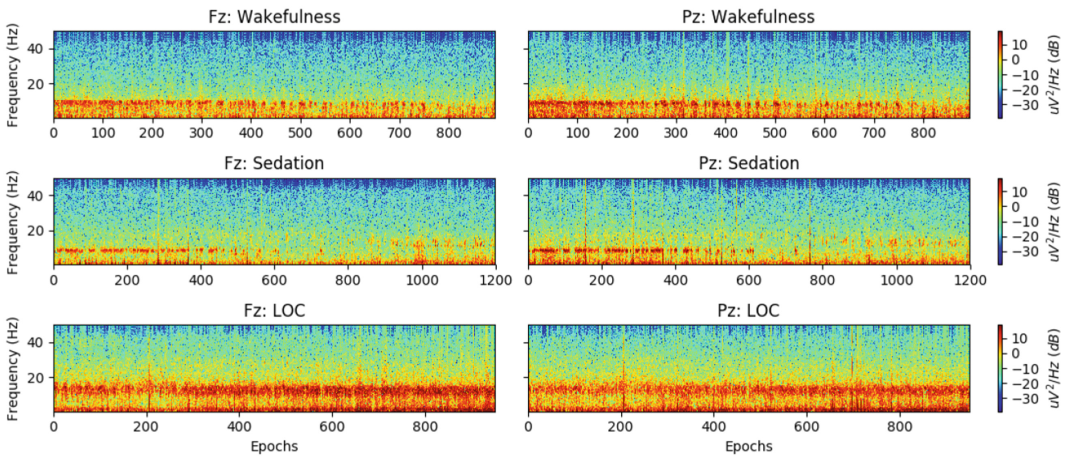

Fig. 6. Power spectral density (uV^{2}/Hz, dB) of the EEG epochs, divided by the sedation phases of the experiment. Representative frontal (Fz) and parietal (Pz) electrodes are shown.

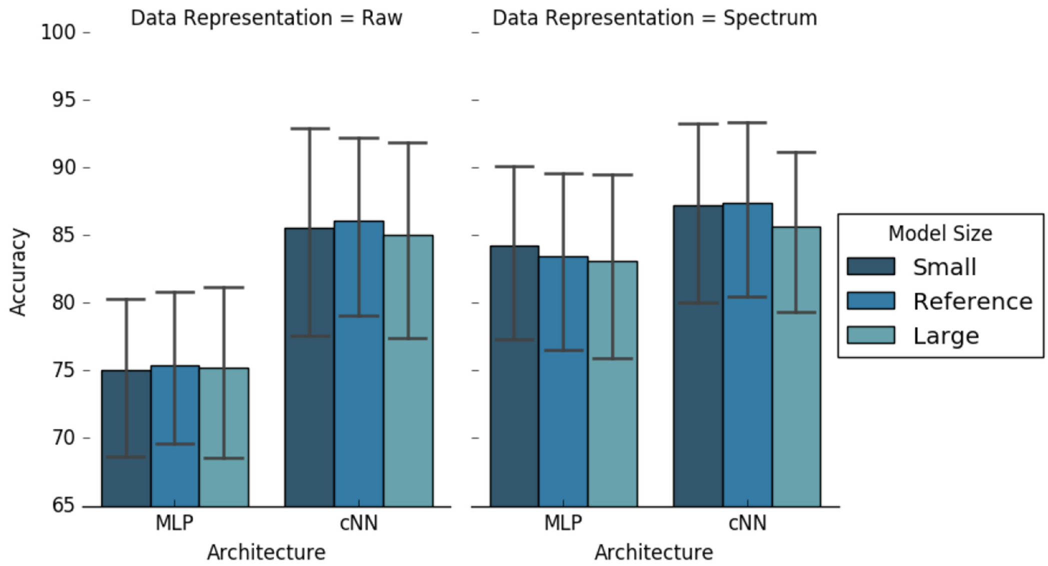

Fig. 7. Three-way ANOVA for the comparison of the data representations, architectures and network sizes. Error bars indicate 95% confidence interval.

Table 2. ANOVA Table

	Sum_sq	df	F	Pr(> F)
Architecture	1230.94	1	10.601	**0.0015**
Data Representation	620.06	1	5.340	**0.0229**
Model Size	14.34	2	0.061	0.9401
Architecture-Data Representation	351.13	1	3.024	0.0852
Architecture-Model Size	6.21	2	0.026	0.9735
Data Representation-Model Size	6.18	2	0.026	0.9737
Architecture-Data Representation-Model Size	0.85	2	0.003	0.9963
Residual	11146.71	96		

representation (Raw/PSD) also had a significant but weaker effect ($F = 5.34$, $p = 0.0229$), driven by the improved accuracy of MLPs with PSD data.

In terms of resource utilization, the cNN was also better than the MLP, as the latter had a significantly larger number of parameters to learn (e.g. 46,872,579 in MLP vs 2,921,219 in cNN, for the reference network size). cNN was also faster to train by $\sim$18%. Furthermore, a repetition of the above experiments with an alternative comparison using the same number of trainable parameters (rather than the same number of neurons) in each architecture, gave a much more prominent difference in accuracies, with the MLP performing much worse. Finally, we also verified that increasing epoch size from 1s through to 10 s did not improve performance of either model.

4 Discussion

Our findings highlight the capability of potential for deep learning of human EEG to discover and utilize generalizable features for automatic identification of consciousness during anesthesia. Further, we have shown that modern cNNs significantly outperform fully connected MLPs, potentially due to their ability to extract more effective spatio-temporo-spectral features from the raw signal. This notion is supported by the fact that MLPs performed as well as cNNs when given PSD data as input.

Though this study aimed to conduct a comparative analysis rather than hyperparameter optimization to maximize accuracy, the fact that cNNs were able to perform very well given only with 1 s of raw EEG data despite the lack of such optimization suggests that they could find utility in real-world applications for assessment and monitoring of consciousness.

Acknowledgements. We acknowledge funding from the UK Engineering and Physical Sciences Research Council [EP/P033199/1], the Belgian National Fund for Scientific Research, the European Commission, the Human Brain Project, the Luminous project, the French Speaking Community Concerted Research Action, the Belgian American Educational Foundation, the Wallonie-Bruxelles Federation, the European Space Agency, the University and University Hospital of Liège (Belgium). This research was undertaken with the support of the Alan Turing Institute (UK Engineering and Physical Sciences Research Council Grant EP/N510129/1).

References

1. Purdon, P.L., et al.: Electroencephalogram signatures of loss and recovery of consciousness from propofol. Proc. Natl. Acad. Sci. (2013). https://doi.org/10.1073/pnas.1221180110
2. Avidan, M.S., et al.: Prevention of intraoperative awareness in a high-risk surgical population. N. Engl. J. Med. (2011). https://doi.org/10.1056/NEJMoa1100403
3. Chennu, S., O'Connor, S., Adapa, R., Menon, D.K., Bekinschtein, T.A.: Brain connectivity dissociates responsiveness from drug exposure during propofol-induced transitions of consciousness. PLoS Comput. Biol. **12**, 1–17 (2016). https://doi.org/10.1371/journal.pcbi.1004669

4. Murphy, M., et al.: Propofol anesthesia and sleep: a high-density EEG study. Sleep **34**, 283–291 (2011). https://doi.org/10.1093/sleep/34.3.283
5. Schirrmeister, R.T., et al.: Deep learning with convolutional neural networks for EEG decoding and visualization. Hum. Brain Mapp. **38**, 5391–5420 (2017). https://doi.org/10.1002/hbm.23730
6. Stober, S., Cameron, D.J., Grahn, J.A.: Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings. Neural Inf. Process. Syst. **2014**, 1–9 (2014)
7. Howbert, J.J., et al.: Forecasting seizures in dogs with naturally occurring epilepsy. PLoS ONE (2014). https://doi.org/10.1371/journal.pone.0081920
8. Park, Y., Luo, L., Parhi, K.K., Netoff, T.: Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. Epilepsia (2011). https://doi.org/10.1111/j.1528-1167.2011.03138.x
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst., 1–9 (2012). http://dx.doi.org/10.1016/j.protcy.2014.09.007
10. Juel, B.E., Romundstad, L., Kolstad, F., Storm, J.F., Larsson, P.G.: Distinguishing anesthetized from awake state in patients: a new approach using one second segments of raw EEG. Front. Hum. Neurosci. (2018). https://doi.org/10.3389/fnhum.2018.00040
11. Korshunova, I., Kindermans, P.-J., Degrave, J., Verhoeven, T., Brinkmann, B.H., Dambre, J.: Towards improved design and evaluation of epileptic seizure predictors. IEEE Trans. Biomed. Eng. **65**, 502 (2017). https://doi.org/10.1109/TBME.2017.2700086