EEG-based Outcome Prediction after Cardiac Arrest with Convolutional Neural Networks: Performance and Visualization of Discriminative Features

Running title: Deep learning for coma-EEG

Authors: Stefan Jonas[1], Andrea O Rossetti[2], Mauro Oddo[3], Simon Jenni[1], Paolo Favaro[1], Frederic Zubler[4,*]

1. Computer Vision Group, Department of Computer Science, University of Bern, Switzerland
2. Department of Neurology, University Hospital (CHUV) & University of Lausanne, Switzerland.
3. Department of Intensive Care Medicine, University Hospital (CHUV) & University of Lausanne, Switzerland.
4. Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Switzerland.

Corresponding Author (*)
Frederic Zubler, MD, PhD
Sleep-Wake-Epilepsy-Center
Department of Neurology
Bern University Hospital - Inselspital Bern
Freiburgstrasse 4
3010 Bern
Switzerland
E-mail: frederic.zubler@gmail.com

Acknowledgments: FZ was supported by the Baasch-Medicus Foundation. The Swiss National Foundations provided financial support to AOR (CR3213_143780) and to MO (32003B_155957). The funding sources had no part in the analysis or decision to publish. The authors thank Jan Novy, MD PhD, Christine Stähli, RN, and Laura Pezzi, RN, for their help in data assessment. None of the authors has a conflict of interest to declare.

Data availability statement: EEG data sharing is not authorized by the local ethical committee.

Abstract

Prognostication for comatose patients after cardiac arrest is a difficult but essential task. Currently, visual interpretation of electroencephalogram (EEG) is one of the main modality used in outcome prediction. There is a growing interest in computer assisted EEG interpretation, either to overcome the possible subjectivity of visual interpretation, or to identify complex features of the EEG signal. We used a 1-dimensional Convolutional Neural Network (CNN) to predict functional outcome based on 19-channel-EEG recorded from 267 adult comatose patients during targeted temperature management after CA. The area under the receiver operating characteristic curve (AUC) on the test set was 0.885. Interestingly, model architecture and fine-tuning only played a marginal role in classification performance. We then used gradient-weighted Class Activation Mapping (Grad-CAM) as visualization technique to identify which EEG features were used by the network to classify an EEG epoch as favorable or unfavorable outcome, and also to understand failures of the network. Grad-CAM showed that the network relied on similar features than classical visual analysis for predicting unfavorable outcome (suppressed background, epileptiform transients). This study confirms that CNNs are promising models for EEG-based prognostication in comatose patients, and that Grad-CAM can provide explanation for the models' decision-making, which is of utmost importance for future use of deep learning models in a clinical setting.

Keywords

deep-learning; convolutional neural networks; interpretability; Grad-CAM; electroencephalogram; coma; hypoxic ischemic encephalopathy; prognostication.

1. Introduction

Prognostication of comatose patients with hypoxic-ischemic encephalopathy (HIE) after cardiac arrest (CA) is one of the most challenging tasks faced by Neurologists and Neuro-intensivists on the Intensive Care Unit [Rossetti et al., 2016]. Early, reliable identification of patients without potential to recover from coma, or with the risk to develop severe neurological disabilities, is of paramount importance in order to inform relatives and to avoid inappropriate continuation of life-supporting treatments [Sandroni et al., 2014]. Currently, prognostication is based on multimodal approaches combining clinical and paraclinical examinations [Rossetti, 2017; Sandroni et al., 2018]. One of the most important modality is the electroencephalogram (EEG) [Thenayan et al., 2010; Hofmeijer et al., 2015; Rossetti et al., 2016; Westhall et al., 2016; Rossetti, 2017; Backman et al., 2018; Sandroni et al., 2018]. In clinical practice, EEG analysis is performed visually by a trained electroencephalographer based on relatively few criteria [Hirsch et al., 2013]. Several EEG patterns have been shown to be associated with either unfavorable outcome (e.g. suppressed background, burst-suppression - especially if showing identical bursts, epileptiform activity) or favorable outcome (continuous background, background reactivity) [Hofmeijer et al., 2015; Westhall et al., 2016]. Quantitative (computer-based) methods have been proposed, either to perform as surrogate electroencephalographers [Tjepkema-Cloostermans et al., 2017], to increase precision and speed of interpretation [Rundgren et al., 2010; Ruijter et al., 2015; Ruijter et al., 2018], or to detect EEG features not easily recognizable by human eye [Beudel et al., 2014; Zubler et al., 2017; Pfeiffer et al., 2018]. Most quantitative approaches are “feature engineered”, meaning that an algorithm was explicitly designed to detect or quantify pre-defined features of the EEG signal such as amplitude, frequency spectrum, presence of spiky elements, and linear or non-linear interactions between the channels [Zubler et al., 2016]. The practical advantage of this approach is also its major limitation, namely that it benefits and depends from the knowledge accumulated over decades by neurophysiologists. However, experienced human interpreters still perform better than specific algorithms, possibly because humans rely on a robustly acquired pattern-recognition ability that is not easy to translate into explicit algorithms [Tjepkema-Cloostermans et al., 2018].

To overcome these limitations, neurophysiologist have begun to use machine-learning techinques, especially deep learning (DL). The practical advantage of DL is that it provides an automated pipeline of “feature extraction” followed by classification, whereby the algorithm itself (and not the programmer) defines which features of the signal are relevant for an accurate classification [LeCun et al., 2015]. In particular, convolutional neural networks (CNNs) - a certain type of DL network loosely inspired by the animal visual system, in which connections between (convolutional) layers are made by sliding filters across the input data - have been demonstrated to be extremely efficient for analyzing images [Krizhevsky et al., 2012]. CNNs and other DL architectures have been applied to EEG data, either for clinical applications such as scoring of sleep stages [Biswal et al., 2018], detection of focal epileptiform discharges [Johansen et al., 2016; Tjepkema-Cloostermans et al., 2018], detection of “abnormality” or “pathologies” in clinical EEGs [Schirrmeister et al., 2017a; van Leeuwen et al., 2019], or for brain-computer-interfaces [Carvalho et al., 2017; Schirrmeister et al., 2017b; Lawhern et al., 2018]. Two recent studies used CNNs for prognostication in comatose patients after CA. In the first, a relatively simple network was applied to raw EEG data and reached very good discriminative performance in predicting the functional outcome of patients [van Putten et al., 2018a]. In the second, different networks were applied to EEG power spectra, with slightly less good performance [Ghassemi, 2018].

While the success of DL for healthcare application seems promising [Faust et al., 2018], one major limitation is likely to hinder its acceptance by clinicians and patients organizations, namely its lack of interpretability [Doshi-Velez and Kim, 2017]. Indeed, it is often difficult to identify which specific features of the input data were influential for a given classification decision. In the last years, several techniques have been developed to explain the decision-making process in CNNs. Some of these have already been applied to EEG-data. One approach is to look at the convolutional filters and their correlation with the different EEG-channels [Lawhern et al., 2018]. Another is to use the network in the reverse direction in order to generate a typical input sample for a given class [van Putten et al., 2018b]. A third approach is to produce a so-called class activation map (CAM, [Zhou et al., 2016]), a sort of “heat map” highlighting regions that support the classification into specific categories. CAMs have been applied to EEG data for instance to assess the effect of exercise on

brain function [Ghosh et al., 2018]. In the context of EEG-based prognostication, this method should allow identifying which features were used by the network to classify an EEG pattern as suggestive for a favorable or unfavorable outcome.

Here, we designed a CNN to predict the clinical outcome in comatose patients after CA based on EEG recorded in the very early phase (during targeted temperature management). The first goal of the present study is to confirm the findings of [van Putten et al., 2018a] on another prospectively acquired cohort of patients. The second and equally important goal is to apply a visualization algorithm to identify the EEG features learnt by the network. For this objective, we use Grad-CAM [Selvaraju et al., 2017], a specific CAM implementation. We demonstrate how the model's classification decisions, the grad-CAM results and the EEG row traces can be combined in order to identify EEG features learned by the network for preforming prognostication. Finally, we analyze different mechanisms leading to misclassification. This last point is especially important in prognostication for comatose patients, as false classifications can have dramatic consequences.

2. Materials and Methods

2.1 Patients and treatment

EEGs were recorded at the University Hospital of Lausanne (CHUV) and were part of prospectively acquired cohort of comatose patients after cardiac arrest (CA). Details of the recruitment and treatment have been described elsewhere [Rossetti et al., 2010; Rossetti et al., 2017]. The study was approved by the ethic commission of Canton of Vaud (Number 16/13); a waiver of consent was granted since EEG is part of the clinical work-up. In this study, we included all comatose patients following resuscitation after CA who survived beyond 24 hours after admission and who underwent an EEG during targeted temperature management (TTM) between September 2012 and April 2018. At the beginning of the recruitment period, TTM consisted of therapeutic hypothermia at 33°C; starting from December 2014 hypothermia was gradually replaced by controlled normothermia at 36°C, which became the standard treatment in June 2016. TTM was started immediately on

admission with ice packs and ice-cold infusions and then with a surface cooling device during 24 hours. During that time, patients were given sedation and analgesia with propofol (4mg/kg/h), midazolam (0.1 mg/kg/h) or fentanyl (1.5 ug/kg/h) and myorelaxant in case of shivering. Decision to withdraw support was taken after at least 72h, based on the presence of at least two of the following: Unreactive EEG background after TTM, treatment-resistant myoclonus or electrographic status epilepticus, bilateral absence of somatosensory-evoked potentials, and absence of at least one of the following brainstem reflexes at 72 hours, off sedation: pupillary, oculocephalic or corneal [Rossetti et al., 2010].

The clinical outcome was prospectively assessed at 3 months using semi-structured interviews with the Cerebral Performance Category (CPC) [Booth et al., 2004]. A CPC value of 1 (no deficit) or 2 (minor deficits) was considered as favorable outcome; a CPC value of 3 (severe deficits), 4 (vegetative state) or 5 (death) was considered an unfavorable outcome.

Patients were randomly split into a training/validation set (80%) and a test set (20%); patients in the training/validation set were then randomly assigned to the training (80%) or the validation (20%) set.

2.2 Data acquisition and presentation

EEG Recordings were performed for 20 minutes with 19 electrodes according to the international 10-20 system, with a reference placed next to Fpz. Initials EEG recordings were at 250 Hz (occasionally 1000 Hz), and were down-sampled to 50 Hz (after low-pass filtering). In addition, we applied a high-pass filter with cutoff frequency of 0.5 Hz to eliminate DC-shifts and low frequency artifacts. The quality of each EEG, especially the presence of electromyographic artifacts, were assessed post-hoc by a certified electroencephalographer (F.Z.) as described in [Zubler et al., 2017].

For each patient we considered the first 5 minutes of EEG without artifacts and in absence of external stimuli. These 5-minute recordings were decomposed into segments of (initially) 10 seconds without overlap called *epochs*, which were presented to the model independently. The input consisted thus of epochs of raw EEG

data presented as a 19 x 500 array (representing the voltage in μV recorded at 19 channels during 10 seconds at 50Hz). However, the input was not considered as 2-dimensional images, but as a 1-dimensional (1-D) image with 19 different "color" channels (at the first convolutional layer, every filter consisted of 19 channel-specific 1-D kernels, which were convolved independently with their corresponding channels before summation, so that already after the first convolutional layer the channel-specific information was mixed). The motivation to use a 1-D representation was to avoid the introduction of an arbitrary neighboring relationship between the different channels (for instance, the channel Fp1 was not *a priori* "closer" to F8 or Fp2 than it was to any other channel).

2.2 Deep learning architecture

We implemented a 1-D convolutional neural network. The architecture was inspired by the VGG network [Simonyan and Zisserman, 2015]. VGG is a widely used architecture, which has shown great performance in image classification. In its original description, VGGNet consists of several 'blocks', each block having 2 to 3 convolutional layers consecutively stacked before max-pooling is applied. At the end, there are two hidden fully connected layers with 4096 neurons, before the final output layer. This architecture is extremely deep and carries an extremely large amount of parameters in the order of fifty million. That makes it at risk of overfitting, especially when the training set is limited, and requires the use of a graphics processing unit for training.

For these reasons, we considered a simplified version of VGG, consisting of a reduced number of 'blocks', each one having two successive convolutional layers before max-pooling, and only one fully-connected hidden layer with fewer neurons. The final output layer consisted of a single sigmoid neuron, representing the probability for unfavorable outcome for the patient from which the EEG epoch was recorded. The model was optimized using the Adam learning algorithm [Kingma and Ba, 2014] minimizing the binary cross-entropy loss function. We refer to this model as t-VGG ("tiny-VGG"). The final model specifications are presented in Table 1.

Each EEG epoch was classified independently, that is, the model assigned a probability for unfavorable clinical outcome to each epoch. The probability for an entire EEG was obtained by averaging the probabilities of all its epochs. We then classified the entire EEG as unfavorable outcome if the average probability reached a given threshold, and as favorable outcome if the averaged probability stayed below that threshold. During model optimization, the threshold was moved to produce a receiver operating characteristic curve (ROC-curve). For the final model evaluation, the threshold was set to 50%.

2.3 Model fine-tuning and hyperparameters

Our model was optimized on the validation set with respect to hyperparameters (segment length and overlap, length and number of filters, number of hidden neurons) and model components (number of layers, regularization). The performance was assessed with the area under the ROC-curve (AUC).

2.4 Training and implementation

Training was performed on a CUDA enabled nVidia GTX. Mini-batches contained 128 EEG segments. Early stopping was performed when training accuracy reached 90% to prevent overfitting.

The models were implemented in Python using the open source deep-learning framework Keras [Chollet and others, 2015] with a TensorFlow backend.

2.5 Visualization

Gradient-weighted class activation mapping (Grad-CAM): To visualize features associated with a specific outcome we implemented the Grad-CAM [Selvaraju et al., 2017] algorithm. Here, a heatmap is created from the last convolutional layer to highlight specific regions of the EEG segments which supported the classification as unfavorable outcome (Class 1). The algorithm was applied to the last convolutional layer, that is, at the highest level in the hierarchy before the temporal aspects of the data are deconstructed in the all-to-all penultimate layer [Selvaraju et al., 2017; Zhou

et al., 2016]. The resolution of the heatmap is determined by the size of the feature maps from the last convolutional layer. As suggested in the original publication, Grad-CAM can also be used to highlight so-called “counterfactual explanations”, namely regions that if removed, could change the network’s classification. We use this approach to highlight regions supportive of a classification as favorable outcome (Class 0).

Global Average Pooling (GAP): Since Grad-CAM is applied to the last convolutional layer, it does not incorporate the computation performed by the last layer (the fully-connected hidden layer), and thus may not be entirely representative of the features supporting the network’s decision. To better identify single regions supporting a decision, we trained a so-called global average pooling (GAP [Zhou et al., 2016]) model, that is, a model similar to our final t-VGG model but lacking the hidden fully-connected layer. The GAP network was used only for our visualization task and was not further optimized. By putting more emphasis on the convolutional layers, more of the overall computation was performed by the network prior to or at the layer probed by Grad-CAM.

2.6 Training with physiological sleep

In an attempt to help the network to better recognize “benign” EEG patterns, we retrained the final t-VGG network from scratch with an additional dataset of 16 sleep EEGs added to the training set (with the label “favorable outcome”). These EEGs were recorded during Non-Rapid-Eye-Movement (NREM) sleep in non-comatose subjects (150 seconds during sleep stage NREM-2 and 150 seconds during NREM-3 per subject).

3. Results

3.1 Patients

303 patients had an EEG during TTM during the recruitment period; 36 patients (12 with favorable outcome) were excluded because of (mainly muscle) artifacts on the EEG, resulting in 267 patients (70 females) being analyzed in the present study. Their

mean age (± SD) was 62.0 (± 14.9) years. The mean latency of EEG recording was 20.3 (± 6.1) hours after CA; 127 patients had favorable (Class 0), and 140 patients unfavorable outcome (Class 1) at three months (of which 118 died). The patients' demographics are shown in Table 2.

3.2 Parameter tuning

Different model parameters were evaluated on the validation set. The two first parameters were the epoch length and overlap. Recordings were split into epochs of various durations (4, 10, 20, or 50 seconds), with and without overlap (0%, 50%, 75%, 90%). The best performance was obtained with an epoch length of 10 seconds with 75% overlap. The other parameters tested were the number of hidden neurons in the penultimate layer, the number of convolutional blocks, the type of pooling, the number of filters in the convolutional layers and filter sizes. For detailed results of the exploration of the other parameters, see Supplementary Material. The total number of parameters of the final model was 16'401.

3.5 Performance of t-VGG

We merged the original training set and the validation set into a final training set containing 213 EEGs, and evaluated the performance of the t-VGG model on the test set consisting of 54 EEGs that were never exposed to the model before. The test set consisted of EEGs from 27 (50%) patients with favorable and 27 patients with unfavorable outcome. The AUC on the test set was 0.885. Detailed results are presented in Table 3.

3.7 Visualization for t-VGG network

The gradient-weighted class activation mapping (Grad-CAM) algorithm for visualization was applied to EEG epochs from the test set during their classification by the optimized t-VGG network. Due to the architecture of the network, the resulting heatmap at the last convolutional layer consisted of 26 datapoints, which defined the temporal resolution for class-discriminative regions for class 1 (unfavorable outcome) and class 0 (favorable outcome). Grad-CAM visualizations for typical EEG epochs

correctly classified are presented in Figure 1 (unfavorable outcome) and 2 (favorable outcome). Grad-CAM visualization for EEG epochs that were misclassified are presented in Figure 3.

Suppressed segments ("flat line") were strongly highlighted by Grad-CAM as class-discriminative regions for unfavorable outcome (Figure 1ab, Figure 3a). In addition, spiky or sharply contoured signals (sharply-contoured generalized periodic pattern, Figure 1b; epileptic spikes, Figure 1c) were often highlighted as supporting the classification for Class 1. By contrast, very few regions were highlighted by Grad-CAM as class-discriminative for favorable outcome - even in epochs that were attributed a very low probability for unfavorable outcome (Figure 2). In several cases, an epoch was classified as favorable outcome without a single region being highlighted as discriminative for this particular class (for instance Figure 2b).

3.8 Visualization for the global averaging pooling (GAP) network

To improve the visualization of class-discriminative features in EEG data, we implemented a global average pooling version of the optimized t-VGG network. As consequence of it lacking one layer, the number of parameters (13'265) of the GAP model was 19% lower than that of the original model. The performance of the GAP model was equal or slightly better than that of the original model on the test set (Table 3). On the validation set, however, performances of the GAP network were slightly lower than that of t-VGG (see Supplementary Material).

The main observation when applying the Grad-CAM algorithm to the GAP network was that many more regions of the EEG epochs classified as class 0 (favorable outcome) were highlighted as discriminative for this class (compare Figure 4a-c with Figure 2a-c). In particular, monomorphic theta rhythms with postero-anterior amplitude gradient were particular class-discriminative for favorable outcome; one representative example is visible in Figure 4b. By contrast, when the visualization algorithm was applied to epochs classified as class 1 (unfavorable outcome) by the GAP network, the results were usually not very different from the visualization applied to the t-VGG network. In particular, flat regions were still discriminative for this class. In a few cases, however, specific local features were more strongly

highlighted by the GAP than by the original model (compare for instance the second discharge in Figure 1b and in Figure 4d).

3.9 Learning physiological sleep patterns

The t-VGG network was then trained on the final training set augmented with examples of physiological non-REM sleep and evaluated on the test set (Table 3). Compared to training without sleep, the AUC was similar, whereas the accuracy for a threshold at 50% probability was slightly reduced. Interestingly, the sensitivity for unfavorable outcome increased whereas the specificity was reduced. We applied Grad-CAM to this new model. In most cases, the visualizations did not show clear differences in discriminative regions between training with and without sleep EEGs When present, delta-waves were sometimes highlighted as discriminative for favorable outcome, whereas sleep-spindles were not.

3.10 Comparison with other deep learning models

We compared the performance of our network on the test set with that of other successful previously published 1-D and 2-D models (adapted to our sampling rate, see Supplementary Material). As first, we implemented a 1-D version of VGG16. This architecture is the “deepest” and has by far the largest number of parameters. The second architecture was the revised EEGNet [Lawhern et al., 2018], a convolutional network specifically developed for EEG-based movement decoding in the context of brain-computer interfaces. It is characterized by special 2-D convolutional operations (such as depthwise or separable convolutions) and was designed to be compact. The third model is DeepConvNet [Schirrmeister et al., 2017b], a 2-D convolutional network used for EEG decoding. The last model is the one proposed by [van Putten et al., 2018a], which was also applied to EEG-based prognostication after CA. It consists of a single convolutional layer (256 feature maps), one hidden layer (128 neurons) and one final output layer. This is the shallowest of all models considered. However, it has more parameters than our model. Detailed results are presented in Table 3. We note that all models showed very similar performance on the test set.

4. Discussion

In this work we implemented a convolutional neural network (t-VGG) for EEG-based prognostication in comatose patients after CA. It consists of three blocks, each containing two sequential convolutional layers and one max-pooling layer, followed by a final hidden fully-connected layer and one decision neuron. Our model was applied to EEG recorded at the early phase, during targeted temperature management, however during a relatively large time window (between 9 and 30 hours after CA). When tested on new data, the AUC of our model was 0.885. This performance is comparable to the one obtained by [van Putten et al., 2018a] on EEGs recorded at a fixed time after CA with a shallower 2-dimensional convolutional network (AUC of 0.89 at 12h after CA, and 0.76 at 24h after CA).

The performance of t-VGG was comparable or exceeded that of other quantitative or clinical approaches previously applied to (smaller, but more homogeneous) parts of the same cohort. For instance, a Bayesian classifier based on 8 pre-determined quantitative EEG features (various bivariate synchronization measures) reached a AUC of 0.81 [Zubler et al., 2017]. A clinical multimodal prognostication approach combining visual EEG features, somatosensory evoked potentials, brainstem reflexes and neuron specific enolase also reached an AUC of 0.81 for predicting unfavorable outcome [Tsetsou et al., 2018]. Our results thus confirm that CNNs applied to raw EEG data are a valuable tool for prognostication after CA.

4.1 Does model architecture matter for EEG data?

We compared our model to other previously published models. Even though the models differ greatly in their architecture specification, such as the number of parameters (ranging from 989 to 42 million), layers (from 1 to 13 convolutional layers) and structure (1-D or 2-D input representation), all achieved relatively good and similar performance. Besides, our model and the one taken from [van Putten et al., 2018a] were developed for this specific task, whereas EEGNet and DeepConvNet have been designed and optimized for brain-computer-interfaces. Furthermore, the increase in performance in our model during architecture selection and fine-tuning

was only 3.1% (AUC) on the validation set. These observations raise the question whether CNN model architecture plays as strong a role for EEG analysis as it does for image processing. The low dependency of model performance on hyperparameter values is reassuring in the optic of future clinical applications, as it suggests good generalization across datasets. Furthermore, as suggested by Schirrmeister et al. [Schirrmeister et al., 2017a], we can postulate that models optimized for one specific EEG-based task could also perform well on other EEG-based classification, which is of advantage, since training data sets are usually sparse. On the other hand, we cannot exclude that we might observe more significant differences between the models with a much larger training data set.

4.2 Lessons from Grad-CAM visualization

We used the Grad-CAM algorithm in order to identify EEG features discriminative for favorable or unfavorable outcome. We recall that these features were not specified ahead of time, but were recognized by the network during training for their association with a specific outcome. Interestingly, the network learned some of the features used by clinicians when visually interpreting an EEG. In particular, suppressed regions, which are often associated with an unfavorable outcome, were recognized as such [Hofmeijer et al., 2015; Westhall et al., 2016]. Also epileptiform transients, which are usually considered as relative markers for unfavorable outcome [Hofmeijer et al., 2015; Westhall et al., 2016] were often discriminative for this class. By contrast, epileptic seizures were not systematically recognized as predictor for unfavorable outcome (for instance Figure 3c) - this is probably due to the limited number of seizures present in the training set, and to the non-stationarity of the EEG signal during a seizure, which makes the learning more difficult.

Very few regions were class-discriminative for favorable outcome in the t-VGG network, even when the epoch was classified as favorable. In a sense, this also bares resemblance with visual analysis, in that EEGs are often considered benign in the absence of malignant features (absence of discontinuity, absence of periodic pattern, absence of epileptiform activity [Hofmeijer et al., 2015; Westhall et al., 2016]). If the absence of malignant feature was a criteria for our network, this cannot be attributed to one specific region. EEG background variability is another EEG feature associated

with favorable outcome [Efthymiou et al., 2017] which can not be attributed to a specific location of the EEG. Background variability might have been used as discriminant feature by the network at the level of the all-to-all layer, however, this point cannot be investigated with the Grad-CAM algorithm (perturbation of the input signal [Becker et al., 2018] could be a complementary approach to test this hypothesis). Finally, background reactivity (that is, modification of the traces in response to a stimulus) is a classical marker for favorable outcome that cannot be recognized by this particular model, because external stimuli were not incorporated to the input data.

It is standard practice to incorporate one or two all-to-all layers at the end of convolutional networks. Since all-to-all layers combine information from the previous layers, the spatial ordering (in the context of time series such as EEG: the temporal ordering) is lost, which explains why the Grad-CAM algorithm is not immediately applicable. Using Grad-CAM at the last convolutional layer provides interesting insights but does not take into account the computation performed higher in the hierarchy. For this reason, we implemented a GAP network to replace the all-to-all layer. Not only does this improve the Grad-CAM visualization, but it also enforces the computation to rely more on local EEG features. As immediate consequence, more EEG regions were highlighted as class-discriminative for favorable outcome. In particular, we often observed that monomorphic theta rhythmic activity with higher amplitude in the posterior channels became class-discriminative for favorable outcome (Figure 3b). Rhythmic theta activity [Synek, 1988] has been previously described as suggestive of a favorable outcome (by contrast, a postero-anterior amplitude gradient has not yet been validated as isolated predictor, even though it has been proposed a condition for “benign” EEG pattern [Westhall et al., 2016]). In addition to facilitating visualization, the overall performance of the GAP network was not inferior than that of the t-VGG.

4.3 Analyzing errors

In order to gain enough confidence in a method to be able to use it in a clinical setting, it is of paramount importance to know its limitations. This includes analyzing in detail, which epochs were misclassified, and trying to understand why. The epoch

displayed in Figure 3a was falsely classified as poor outcome, whereas the patient survived. Grad-CAM confirmed that mainly the suppressed regions were supporting the incorrect classification. The reason for the error becomes clear when we note that the EEG was recorded very early (13h after CA). It is well known that in the very acute phase or under sedation a burst-suppression pattern is not always associated with an unfavorable outcome [Caporro et al., 2019; Cloostermans et al., 2012]. Using EEG recorded at a later stage might help solve this particular problem. The false positive error presented in Figure 3b is due to another mechanism: an experienced encephalographer would have recognized transients with triphasic appearence, which are potentially caused by metabolic disturbances and not due to a severe hypoxic/anoxic encephalopathy. There were very few triphasic transients on the training set (this pattern appears rarely in the acute phase), which explains the misclassification.

Some of the false negatives can also be explained by looking at the raw EEG data: The EEG epoch represented in Fig 3c consists of (part of) an epileptic seizure. As stated previously, seizures were rare in our collective in the early phase, so that the network had few examples to learn this pattern and associate it with a poor outcome. As for the EEG reproduced in Fig 3d, it would probably have been classified a favorable outcome by a clinician as well. Accordingly, the patient first regained consciousness, but died 7 days later after transfer to another hospital.

4.4 Influencing learning

When interpreting the EEG of a comatose patient, clinicians do not only rely on previously seen coma-EEG, but also evoke their longtime acquired experience with EEG recordings in other conditions. In the context of prognostication after CA, clinicians might for instance recognize the presence of typical patterns of non-rapid eye movement (NREM) sleep, which can be suggestive of a favorable outcome for patients in the Intensive Care Unit [Murray et al., 2009; Sandsmark et al., 2016]. We wanted to determine whether the performance of the network could be improved by teaching it to recognize elements of sleep stage NREM-2 (sleep-spindles or K-complexes) or NREM 3 (rhythmic delta activity). Our motivation was to increase the specificity for unfavorable outcome, since false positive errors have more dramatic

consequences than false negative errors when it comes to deciding on withdrawal of life-supporting treatment. However, the opposite occurred: The sensitivity for unfavorable outcome increased, whereas the specificity decreased. One can postulate that training the network with EEG epochs from non-comatose subjects introduced a bias in the model's representation of favorable coma-EEGs. Grad-CAM showed that after training with sleep, delta-wave were more highlighted as class-discriminative for favorable outcome, whereas it was not the case for sleep-spindles. One possible explanation for this discrepancy is the different frequency of these elements (NREM sleep contains numerically more delta waves than spindles). However, changes in the class activation maps did not always correlate with changes in the classification.

4.5 Strengths and Limitations

In this paper, we used data from a well-characterized and prospectively acquired cohort. We used EEG obtained during the early phase (during TTM), because these recordings were not taken into account for decision to withdraw life-supporting treatment. As such, the risk of self-fulfilling prophecy appears low. We analyzed and presented in detail the correlation between the model decisions, the visualization and the raw EEG data – a step that is unfortunately often neglected, but of importance to investigate the capabilities and limitations of deep learning models in a clinical setting.

Using a 1-D Convolutional Neural Network instead of a 2-D does not seem to affect performance (Table 3). However, this restricts the use of Grad-CAM to the temporal dimension, and does not allow channel-specific feature visualization. This limitation might be only of mild relevance for coma EEG, since the patterns are usually homogenous, at least symmetrical. For other applications, such as focal epileptic activity, a 2-D visualization might be necessary. The variable time delay at which EEG was recorded is another potential limitation, since EEG-patterns can vary in the first 30h. On the other hand, the fact that our model performs well despite a relatively large time-window of recording gives us confidence for an application in the real-world. An additional issue is the size of the dataset: Deep Learning models can usually reach their best accuracies with very large datasets. Whether this also applies to EEG-based classifications and whether the performance would improve with more

training samples, remains to be investigated. Finally, our model incorporates EEG data and not other modalities, such as clinical information, somatosensory evoked potentials, biological markers and MRI findings. Future implementations could include such data, for instance as additional unit in the fully connected layer [van Leeuwen et al., 2019].

5. Conclusion

Deep Learning (DL) has proven to be very successful for classification tasks in a wide range of applications. There is growing evidence that DL can be potentially useful for clinical applications. However, DL has not been yet incorporated into standard decision-making procedures. Our results show that class activation maps can provide visual interpretations for classification based on time-series such as EEG. It is our opinion that future work on DL and EEG should not only aim at improving performance, but also model interpretability. This last point will prove essential if DL is to be entrusted in the near future for contributing to clinical decision. Moreover, gaining a better understanding of the type of patterns used by a network when evaluating an EEG might help us to identify new types of patterns, which might later even be included into visual analysis.

References

Backman S, Cronberg T, Friberg H, Ullén S, Horn J, Kjaergaard J, Hassager C, Wanscher M, Nielsen N, Westhall E (2018): Highly malignant routine EEG predicts poor prognosis after cardiac arrest in the Target Temperature Management trial. Resuscitation 131:24–28.

Becker SL, Ackermann M, Lapuschkin S, Müller K-R, Samek W (2018): Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. CoRR abs/1807.03418.

Beudel M, Tjepkema-Cloostermans MC, Boersma JH, van Putten MJAM (2014): Small-World Characteristics of EEG Patterns in Post-Anoxic Encephalopathy. Front Neurol 5. http://journal.frontiersin.org/article/10.3389/fneur.2014.00097/abstract.

Biswal S, Sun H, Goparaju B, Westover MB, Sun J, Bianchi MT (2018): Expert-level sleep scoring with deep neural networks. J Am Med Inform Assoc JAMIA 25:1643–1650.

Booth CM, Boone RH, Tomlinson G, Detsky AS (2004): Is This Patient Dead, Vegetative, or Severely Neurologically Impaired?: Assessing Outcome for Comatose Survivors of Cardiac Arrest. JAMA 291:870.

Caporro M, Rossetti AO, Seiler A, Kustermann T, Nguepnjo Nguissi NA, Pfeiffer C, Zimmermann R, Haenggi M, Oddo M, De Lucia M, Zubler F (2019): Electromyographic reactivity measured with scalp-EEG contributes to prognostication after cardiac arrest. Resuscitation 138:146–152.

Carvalho SR, Filho IC, Resende DOD, Siravenha AC, Souza CD, Debarba HG, Gomes B, Boulic R (2017): A Deep Learning Approach for Classification of Reaching Targets from EEG Images. In: . 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) pp 178–184.

Chollet F, others (2015): Keras. https://keras.io.

Cloostermans MC, van Meulen FB, Eertman CJ, Hom HW, van Putten MJAM (2012): Continuous electroencephalography monitoring for early prediction of neurological outcome in postanoxic patients after cardiac arrest: a prospective cohort study. Crit Care Med 40:2867–2875.

Doshi-Velez F, Kim B (2017): Towards A Rigorous Science of Interpretable Machine Learning. In: .

Efthymiou E, Renzel R, Baumann CR, Poryazova R, Imbach LL (2017): Predictive value of EEG in postanoxic encephalopathy: A quantitative model-based approach. Resuscitation 119:27–32.

Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR (2018): Deep learning for healthcare applications based on physiological signals: A review. Comput Methods Programs Biomed 161:1–13.

Ghassemi M (2018): Life After Death: Techniques for the Prognostication of Coma Outcomes after Cardiac Arrest. PhD Thesis; Massachusetts Institute of Technology.

Ghosh A, dal Maso F, Roig M, Mitsis GD, Boudrias M-H (2018): Deep Semantic Architecture with discriminative feature visualization for neuroimage analysis. ArXiv E-Prints:arXiv:1805.11704.

Hirsch LJ, LaRoche SM, Gaspard N, Gerard E, Svoronos A, Herman ST, Mani R, Arif H, Jette N, Minazad Y, Kerrigan JF, Vespa P, Hantus S, Claassen J, Young GB, So E, Kaplan PW, Nuwer MR, Fountain NB, Drislane FW (2013): American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology: 2012 version. J Clin Neurophysiol 30:1–27.

Hofmeijer J, Beernink TMJ, Bosch FH, Beishuizen A, Tjepkema-Cloostermans MC, van Putten MJAM (2015): Early EEG contributes to multimodal outcome prediction of postanoxic coma. Neurology 85:137–143.

Johansen AR, Jin J, Maszczyk T, Dauwels J, Cash SS, Westover MB (2016): Epileptiform spike detection via convolutional neural networks. Proc IEEE Int Conf Acoust Speech Signal Process ICASSP Conf 2016:754–758.

Kingma D, Ba J (2014): Adam: A Method for Stochastic Optimization. Int Conf Learn Represent.

Krizhevsky A, Sutskever I, Hinton GE (2012): ImageNet Classification with Deep Convolutional Neural Networks. In: . Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. USA: Curran Associates Inc. NIPS'12 pp 1097–1105. http://dl.acm.org/citation.cfm?id=2999134.2999257.

Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ (2018): EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. J Neural Eng 15:056013.
LeCun Y, Bengio Y, Hinton G (2015): Deep learning. Nature 521:436–444.
van Leeuwen KG, Sun H, Tabaeizadeh M, Struck AF, van Putten MJ a. M, Westover MB (2019): Detecting abnormal electroencephalograms using deep convolutional networks. Clin Neurophysiol Off J Int Fed Clin Neurophysiol 130:77–84.
Murray DM, Boylan GB, Ryan CA, Connolly S (2009): Early EEG findings in hypoxic-ischemic encephalopathy predict outcomes at 2 years. Pediatrics 124:e459–467.
Pfeiffer C, Nguissi NAN, Chytiris M, Bidlingmeyer P, Haenggi M, Kurmann R, Zubler F, Accolla E, Viceic D, Rusca M, Oddo M, Rossetti AO, De Lucia M (2018): Somatosensory and auditory deviance detection for outcome prediction during postanoxic coma. Ann Clin Transl Neurol 5:1016–1024.
van Putten MJAM, Hofmeijer J, Ruijter BJ, Tjepkema-Cloostermans MC (2018a): Deep Learning for outcome prediction of postanoxic coma. In: Eskola, H, Väisänen, O, Viik, J, Hyttinen, J, editors. EMBEC & NBC 2017. Singapore: Springer Singapore. pp 506–509.
van Putten MJAM, Olbrich S, Arns M (2018b): Predicting sex from brain rhythms with deep learning. Sci Rep 8:3069.
Rossetti AO (2017): Clinical neurophysiology for neurological prognostication of comatose patients after cardiac arrest. Clin Neurophysiol Pract 2:76–80.
Rossetti AO, Oddo M, Logroscino G, Kaplan PW (2010): Prognostication after cardiac arrest and hypothermia: A prospective study. Ann Neurol 67:301–307.
Rossetti AO, Rabinstein AA, Oddo M (2016): Neurological prognostication of outcome in patients in coma after cardiac arrest. Lancet Neurol 15:597–609.
Rossetti AO, Tovar Quiroga DF, Juan E, Novy J, White RD, Ben-Hamouda N, Britton JW, Oddo M, Rabinstein AA (2017): Electroencephalography Predicts Poor and Good Outcomes After Cardiac Arrest: A Two-Center Study. Crit Care Med 45:e674–e682.
Ruijter BJ, Hofmeijer J, Tjepkema-Cloostermans MC, van Putten MJAM (2018): The prognostic value of discontinuous EEG patterns in postanoxic coma. Clin Neurophysiol Off J Int Fed Clin Neurophysiol 129:1534–1543.
Ruijter BJ, van Putten MJAM, Hofmeijer J (2015): Generalized epileptiform discharges in postanoxic encephalopathy: Quantitative characterization in relation to outcome. Epilepsia 56:1845–1854.
Rundgren M, Westhall E, Cronberg T, Rosén I, Friberg H (2010): Continuous amplitude-integrated electroencephalogram predicts outcome in hypothermia-treated cardiac arrest patients: Crit Care Med 38:1838–1844.
Sandroni C, Cariou A, Cavallaro F, Cronberg T, Friberg H, Hoedemaekers C, Horn J, Nolan JP, Rossetti AO, Soar J (2014): Prognostication in comatose survivors of cardiac arrest: An advisory statement from the European Resuscitation Council and the European Society of Intensive Care Medicine. Intensive Care Med 40:1816–1831.
Sandroni C, D'Arrigo S, Nolan JP (2018): Prognostication after cardiac arrest. Crit Care Lond Engl 22:150.
Sandsmark DK, Kumar MA, Woodward CS, Schmitt SE, Park S, Lim MM (2016): Sleep Features on Continuous Electroencephalography Predict Rehabilitation

Outcomes After Severe Traumatic Brain Injury. J Head Trauma Rehabil 31:101–107.

Schirrmeister RT, Gemein L, Eggensperger K, Hutter F, Ball T (2017a): Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. 2017 IEEE Signal Process Med Biol Symp SPMB:1–7.

Schirrmeister RT, Springenberg JT, Fiederer LDJ, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W, Ball T (2017b): Deep learning with convolutional neural networks for EEG decoding and visualization. Hum Brain Mapp 38:5391–5420.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017): Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: . 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE. pp 618–626. http://ieeexplore.ieee.org/document/8237336/.

Simonyan K, Zisserman A (2015): Very Deep Convolutional Networks for Large-Scale Image Recognition. In: . International Conference on Learning Representations.

Synek VM (1988): Prognostically important EEG coma patterns in diffuse anoxic and traumatic encephalopathies in adults. J Clin Neurophysiol Off Publ Am Electroencephalogr Soc 5:161–174.

Thenayan EAL, Savard M, Sharpe MD, Norton L, Young B (2010): Electroencephalogram for prognosis after cardiac arrest. J Crit Care 25:300–304.

Tjepkema-Cloostermans MC, de Carvalho RCV, van Putten MJAM (2018): Deep learning for detection of focal epileptiform discharges from scalp EEG recordings. Clin Neurophysiol Off J Int Fed Clin Neurophysiol 129:2191–2196.

Tjepkema-Cloostermans MC, Hofmeijer J, Beishuizen A, Hom HW, Blans MJ, Bosch FH, van Putten MJAM (2017): Cerebral Recovery Index: Reliable Help for Prediction of Neurologic Outcome After Cardiac Arrest. Crit Care Med 45:e789–e797.

Tsetsou S, Novy J, Pfeiffer C, Oddo M, Rossetti AO (2018): Multimodal Outcome Prognostication After Cardiac Arrest and Targeted Temperature Management: Analysis at 36 °C. Neurocrit Care 28:104–109.

Westhall E, Rossetti AO, van Rootselaar A-F, Wesenberg Kjaer T, Horn J, Ullén S, Friberg H, Nielsen N, Rosén I, Åneman A, Erlinge D, Gasche Y, Hassager C, Hovdenes J, Kjaergaard J, Kuiper M, Pellis T, Stammet P, Wanscher M, Wetterslev J, Wise MP (2016): Standardized EEG interpretation accurately predicts prognosis after cardiac arrest. Neurology 86:1482–1490.

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016): Learning Deep Features for Discriminative Localization. In: . 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 2921–2929.

Zubler F, Bandarabadi M, Kurmann, Rebekka, Gast H, Schindler K (2016): Quantitative EEG in the Intensive Care Unit. Epileptologie 33:166–172.

Zubler F, Steimer A, Kurmann R, Bandarabadi M, Novy J, Gast H, Oddo M, Schindler K, Rossetti AO (2017): EEG synchronization measures are early outcome predictors in comatose patients after cardiac arrest. Clin Neurophysiol 128:635–642.

Tables

Block	Layer	Description
1	Input	1-D input of 10s long EEG epochs
	1-D Convolution	16 filters, kernel length 3, stride 1
	BatchNormalization	
	Activation	ReLU
	1-D Convolution	16 filters, kernel length 3, stride 1
	BatchNormalization	
	Activation	ReLU
	Max-Pooling	pool size 4, stride 4
2	1-D Convolution	32 filters, kernel length 3, stride 1
	BatchNormalization	
	Activation	ReLU
	1-D Convolution	32 filters, kernel length 3, stride 1
	BatchNormalization	
	Activation	ReLU
	Max-Pooling	pool size 4, stride 4
3	1-D Convolution	32 filters, kernel length 3, stride 1
	BatchNormalization	
	Activation	ReLU
	1-D Convolution	32 filters, kernel length 3, stride 1
	BatchNormalization	
	Activation	ReLU activation
	Max-Pooling	pool size 4, stride 4
Classification	Flatten	
	Dense	16 neurons, L2 regularization λ=0.01
	BatchNormalization	
	Activation	ReLU
	Dropout	drop probability 50%
	Dense	1 output neuron
	Activation	Sigmoid

Table 1: t-VGG architecture. The 1-D model consists of 3 blocks, each containing 2 sequential convolutional layers before max-pooling is applied. The output layer is a single sigmoid neuron representing the probability of unfavorable outcome. The implementation code is provided in the Supplementary Material.

	Favorable outcome	Unfavorable outcome	p
N	127	140	n.a.
Female	30 (24%)	40 (29%)	0.358
Age (±SD) [y]	59.5 (± 15.1)	64.3 (±14.4)	0.011
Cardiac etiology	114 (90%)	90 (64%)	<0.001
Asystole or pulseless electrical activity on site	26 (20%)	87 (62%)	<0.001
Therapeutic hypothermia (33° C)	54 (43%)	37 (26%)	0.006
Latency of EEG recording (±SD) [h]	19.7 (±5.4)	20.9 (±7.6)	0.11
Disontinuous EEG background	45 (0.35%)	18 (19%)	< 0.001
Areactive EEG background	8 (6%)	84 (60%)	< 0.001
Irritative EEG	2 (2%)	44 (31%)	< 0.001
Patients sedated with propofol	41 (32%)	33 (24%)	0.112
Propofol dosis (±SD) [mg/kg/h]	1.92 (±1.1)	1.97 (±1.2)	0.961
Patients sedated with midazolam	67 (53%)	42 (30%)	<0.001
Midazolam dosis (±SD) [mg/kg/h]	0.12 (±0.05)	0.12 (±0.04)	0.806
Patients sedated with fentanyl	40 (31%)	29 (21%)	0.044
Fentanyl dosis(±SD) [mg/kg/h]	1.38 (±0.9)	1.03 (±0.57)	0.069

Table 2: Patients demographics. Differences between groups were assessed with Mann-Whitney-U tests for numerical values and with Chi-square tests for categorical data.

Model	AUC [95% CI]	EEG Epoch Accuracy	Entire EEG Accuracy	Sensitivity	Specificity	Parameters
t-VGG	0.885 [0.779 - 0.964]	78.73%	83.33%	77.77%	88.88%	16'401
t-VGG GAP	0.900 [0.805 - 0.976]	79.78%	87.04%	85.18%	88.88%	13'265
t-VGG trained with Sleep	0.881 [0.773 - 0.963]	77.73%	81.48%	81.48%	81.48%	16'401
1D-VGG	0.880 [0.782 - 0.963]	75.65%	81.48%	81.48%	81.48%	42'721'473
EEGNet, adjusted	0.849 [0.733 - 0.947]	76.74%	79.63%	74.07%	85.18%	989
DeepConvNet, adjusted	0.866 [0.758 - 0.953]	78.79%	79.63%	74.07%	85.18%	98'401
[van Putten et al., 2018a], adjusted	0.857 [0.738 - 0.954]	76.11%	79.63%	77.77%	81.48%	1'225'729

Table 3: Final performance. Performance of the model(s) trained on the training and validation set, and tested on the 54 patients of the test set. AUC : area under the receiver operating characteristic curve for whole EEGs. Epoch accuracy: percentage of EEG epochs (10-second segments) predicting the correct clinical outcome (threshold for predicting unfavorable outcome was set to 50%); Entire EEG accuracy: percentage of EEGs predicting the correct outcome based on averaged probability of all epochs it contains (threshold set to 50%). The sensitivity and specificity are toward prediction of unfavorable outcome. For description of the models see text.

Figure Legends

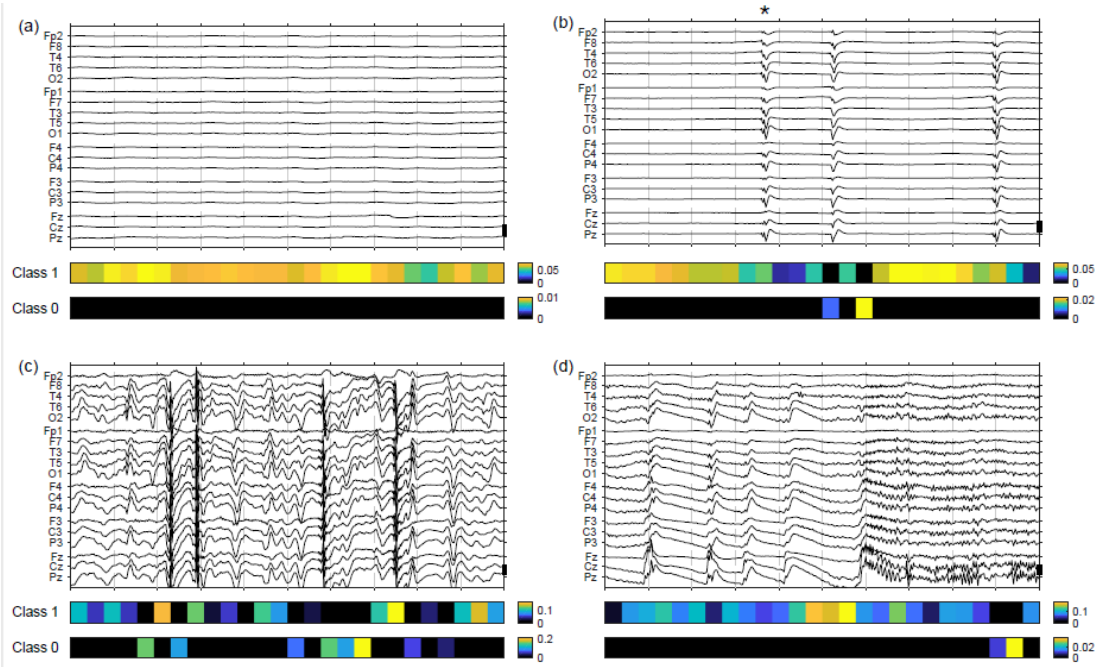

Figure 1: Grad-CAM visual explanation for EEG epochs correctly classified as unfavorable outcome (true positives for unfavorable outcomes) by the t-VGG network. For each example, the figure displays a 10-second EEG epoch in a pseudo-monopolar montage (vertical thick bar = 100 uV) together with its corresponding class activation maps for unfavorable outcome (Class 1) and favorable outcome (Class 0). Each map contains 26 data points, corresponding to 26 temporal regions of the EEG epoch (duration of one region = 385 ms). The lighter the color, the stronger a region is discriminative for a particular class. **(a)** EEG of a 75y female, treated with controlled normothermia (CNT), recorded 22 h after CA, with unfavorable outcome (CPC 5). The probability for unfavorable outcome attributed by the network was maximal (P=1.0). The EEG showed a continuously suppressed background without superimposed periodic pattern. Most of the temporal regions were marked as supporting the classification for unfavorable outcome. No temporal region was supporting the classification for favorable outcome. **(b)** EEG of a 65y male, CNT, recorded 10h after CA, CPC 5, P =1.0: The EEG showed a suppressed background with superimposed sharply-contoured pseudo-periodic generalized discharges. The suppressed segments longer than 2 seconds, and to a lesser extent two of the generalized discharges (the ones with a “poly-spike” configuration,*) were class-discriminative for unfavorable outcome; the third discharge (°) and an after coming segment were discriminative for favorable outcome. **(c)** EEG of a 70y female, CNT, recorded 21h after CA, CPC 5, P = 0.99: The EEG showed a delta/theta continuous background with abundant epileptic spikes. The majority (*), but not all, of the spikes were within or next to regions strongly discriminative for unfavorable outcome. **(d)** EEG of a 44y female, CNT, recorded 20.5 h after CA, CPC 5, P = 0.60: The EEG epoch showed initially interictal epileptic discharges, and after 6 seconds the begin of an electroencephalographic seizure. Mainly the “descending” segments of discharges, with superimposed fast activity, supported the classification for unfavorable outcome.

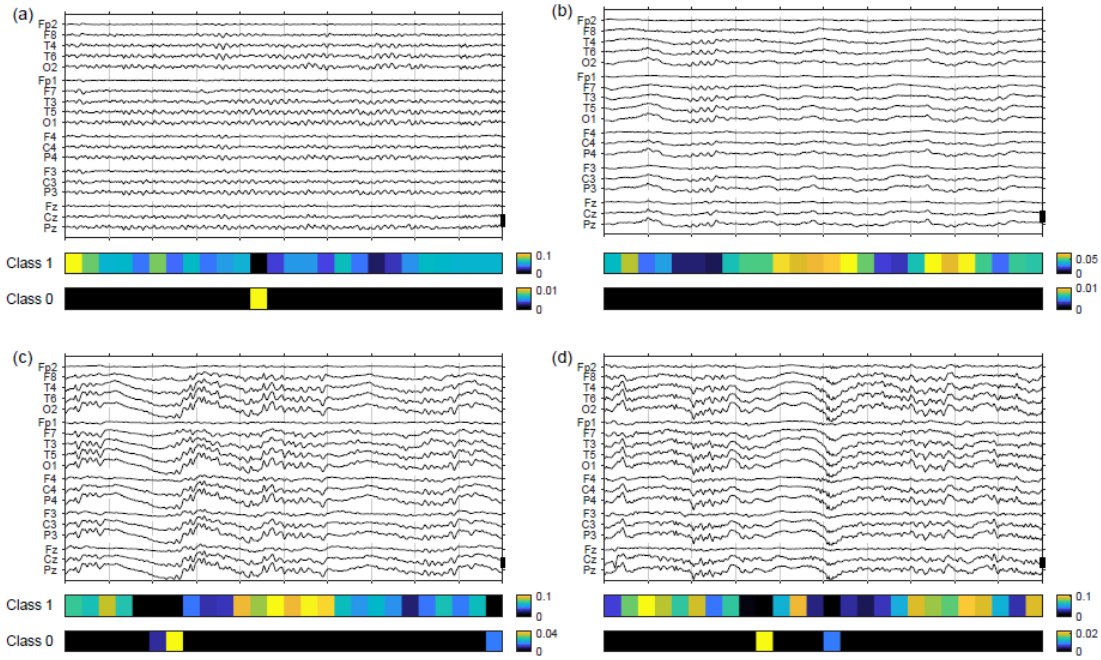

Figure 2: Grad-CAM visual explanation for EEG epochs correctly classified as favorable outcome (true negatives for unfavorable outcome) by the t-VGG network. For interpretation of the color bars see caption of Figure 1. Despite the EEG epochs being correctly recognized as a favorable pattern, only very few of the single temporal regions were marked by the Grad-CAM algorithm as class-discriminative for favorable outcome. **(a)** EEG of a 85y male, treated with controlled normothermia (CNT), recorded 24h after CA, who recovered without deficits (CPC 1). The probability attributed by the network for an unfavorable outcome was low (P = 0.08). The EEG showed a continuous theta background with physiological antero-posterior gradient. **(b)** 82y male, CNT, recorded 22h after CA, CPC 1, P =0.08. Delta-background with superposition of spindle-shaped theta-activity. **(c)** EEG of a 46y female, treated with therapeutic hypothermia, recorded 24h after CA, CPC 1, P=0.01. Diffuse theta and alpha activity modulated by single delta-wave (K-complex-like). **(d)** EEG of a 65y male, CNT, recorded 20h after CA, CPC 1, P = 0.08: Theta and alpha activity with isolated delta-waves followed by faster activity.

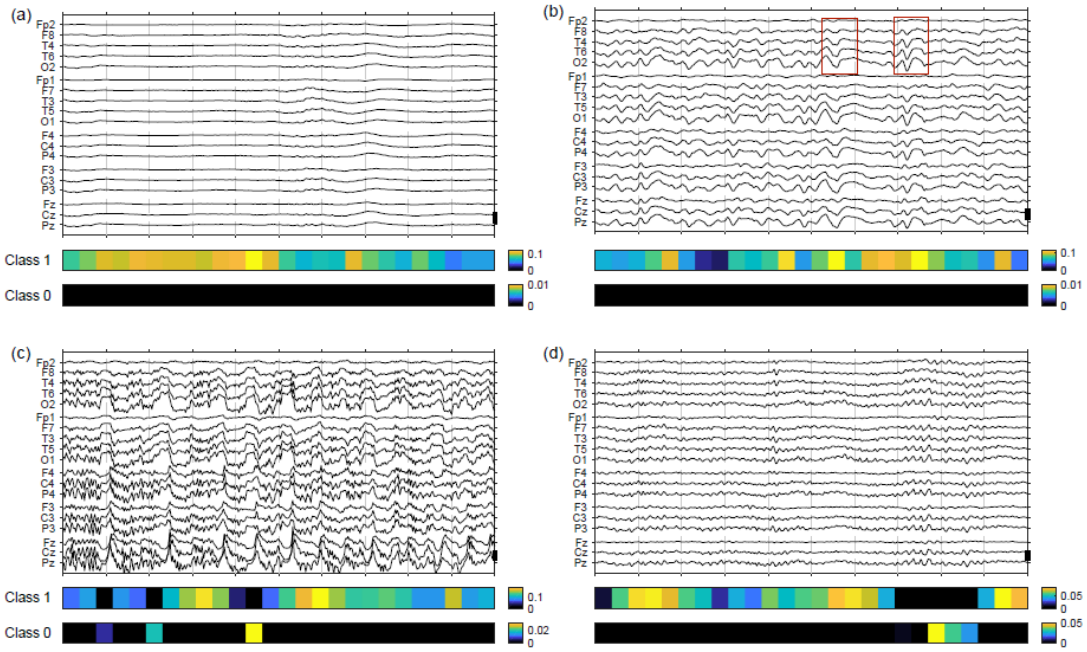

Figure 3: Grad-CAM visual explanation for EEG epochs that were misclassified by the t-VGG network (a, b: false positives; c, d: false negatives for unfavorable outcome). For interpretation of the color bars see caption of Figure 1. **(a)** EEG of a 56y male, treated with therapeutic hypothermia, recorded 13h after CA. The patient recovered without deficits (CPC1), however the network attributed a high probability for unfavorable outcome (P=0.98): The EEG shows a burst-suppression (with not highly epileptiform bursts). Suppression regions were class-discriminative for unfavorable outcome. **(b)** EEG of a 65y male, treated with controlled normothermia (CNT), recorded 30h after CA, who recovered with minor deficits (CPC 2). The network attributed a slightly higher probability for unfavorable outcome (P=0.58). On the EEG we observed irregular theta background with so-called triphasic waves, a typical finding in case of metabolic encephalopathy (boxes). Slower signal components (delta waves) were discriminative for unfavorable outcome. **(c)** Same patient than in Figure 1c, other EEG epoch. Despite the patient having an unfavorable outcome, the probability attributed for unfavorable outcome was relatively low (P = 0.26). This EEG epoch was registered at the end of an epileptic seizure. **(d)** 57y female, CNT, recorded 22h after CA, CPC 5, P = 0.01. The EEG showed a variable and continuous theta/alpha background, suggesting a mild hypoxic-anoxic encephalopathy. The patient indeed awoke, but died one week later.

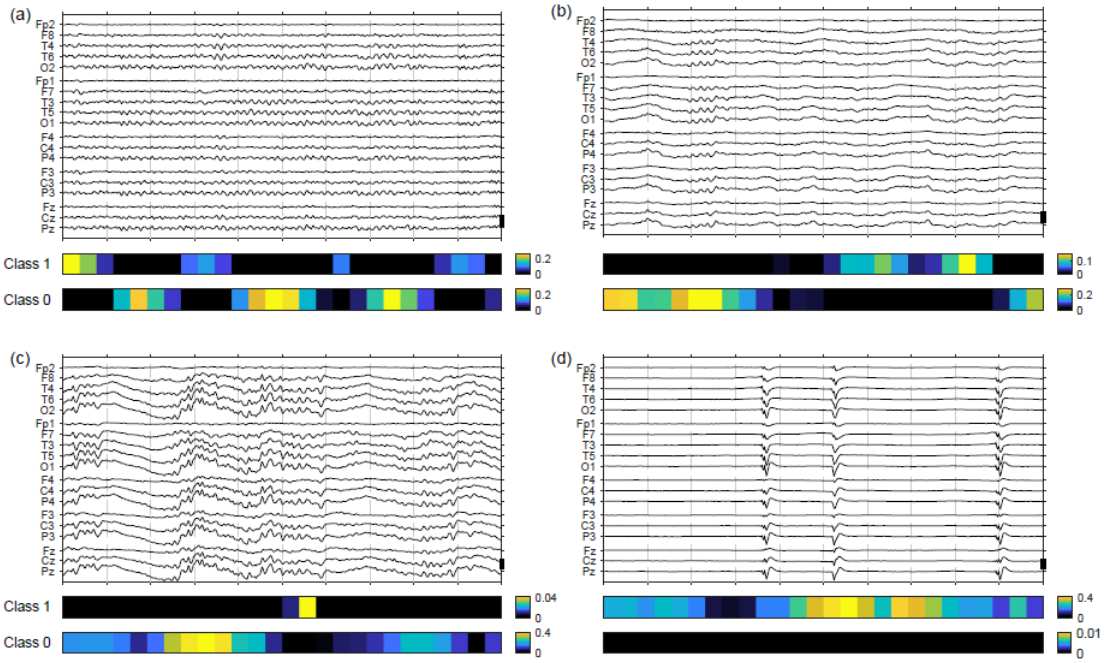

Figure 4: Visual explanation for EEG epochs correctly classified by the global averaging pooling (GAP) network. For interpretation of the color bars see caption of Figure 1. **(a, b, c)** Same EEG epochs than in Figures 2abc. The GAP model correctly attributed low probability for unfavorable outcome to these three epochs (probability of 0.02, 0.06, and 0.01 respectively). Much more temporal regions were class-discriminative for favorable outcome than for the model with all-to-all penultimate layer, in particular segments with monomorphic theta rhythms with postero-anterior amplitude gradient (see boxes; signals from the right temporal electrode chain are included for illustrative purposes only). **(d)** Same EEG epoch as in Figure 1d. The probability for unfavorable outcome was correctly estimated as being very high (P = 1.0). With the GAP architecture, the second sharply configured discharge (°) and its immediate surrounding became the most discriminative regions for unfavorable outcome.