TOWARDS AN AUTOMATED ANALYSIS FOR THE DELINEATION OF FOCAL EPILEPSY

Single Pulse Electrical Stimulation and Machine Learning

Master Thesis Technical Medicine

By

Emile d'Angremont

April, 2018

Graduation committee:
M.J.A.M. van Putten, MD, PhD
F.S.S. Leijten, MD, PhD
R.J. Haarman, MSc
G.J.M. Huiskamp, PhD
A. Garde Martinez, PhD

UNIVERSITY OF TWENTE.

UMC Utrecht

"For as in ordinary life every person's disposition, and the concealed feelings of the mind and passions are most drawn out when they are disturbed – so the secrets of nature betray themselves more readily when tormented by art than when left to their own course."

— Sir Francis Bacon, *Novum Organum (1620)*

Preface

Here it is. The result of nearly eight years of study summarized in a thesis of barely fifty pages. OK to be fair, the content is only about my graduation project, but with it, I hope to show my development from a first year student to the clinical physician that I am about to become.

In the past 11 months, I did my final internship at the Functional Neurosurgery and Epilepsy department of the UMC Utrecht. It truly was an inspiring experience and I enjoyed every part of it. On the one hand, I worked really hard on developing myself in the field of machine learning and applying it to SPES data, but on the other hand, I could clinically develop myself in a department that very quickly made me feel at home. Of course, I need to thank a number of people for that.

First of all, Frans Leijten, thank you for sharing some of your endless knowledge. Your lectures and clinical supervision were truly inspiring. I think every medical specialism needs a doctor like you. Geertjan Huiskamp, thank you for being there whenever you were needed, which happened almost daily, to answer all my questions and to help me out when I got stuck. I think every research department needs an engineer like you. Rian Haarman, I think you helped me gaining some interesting insights about myself, but you also were a strong motivational factor to me. I always felt like I could conquer the world after a meeting with you. Michel van Putten, I think in many ways you are an example to me and many other technical medicine students. Thanks for being my technological supervisor and the chairman of my graduation committee. Christoph Brune, thank you for the very helpful advice on machine learning. Ainara Garde Martinez, thanks for being the external member of my graduation committee.

Daniël and Michelle, but also Banu, Matteo, Jurgen, Nicole, Dorien and Willemiek, I really enjoyed your company in our 'student room' or during lunch. You provided exactly the kind of distraction that I needed. The same goes for Cyrille Ferrier, Maeike Zijlmans, Sandra van der Salm and Tineke Gebbink; I think you all are of great value to the department and I enjoyed every interaction with you.

Summary

In patients with drug-resistant focal epilepsy, surgery can be considered. The goal is to remove the epileptogenic tissue, while sparing the eloquent cortex. Prior to surgery, a prolonged electroencephalography (ECoG) recording can assist in the delineation of epileptogenic tissue and functionality of the surrounding cortex. During these recordings, single pulse electrical stimulation (SPES) of the intra-cranial electrodes is performed to evoke pathological responses from the epileptogenic tissue, which occur >100 ms after stimulation. These responses are called delayed responses (DRs). In the UMC Utrecht, they are visually analyzed by use of time-frequency (TF-SPES) images from approximately 2 sec. around stimulation. Each image is scored by two human observers on the presence of an evoked DR in three different frequency bands, namely spikes (10-80 Hz), ripples (80-250 Hz) and fast ripples (250-520 Hz). This visual analysis is very labor intensive. An additional problem is that DRs are occasionally observed as a physiological phenomenon.

In the first part of this research, we trained a support vector machine (SVM) and a convolutional neural network (CNN) with the aim to automatically detect and classify the DRs in TF-SPES images. The training data consisted of 47197 images from 15 patients, with the consensus of two human observers as ground truth. The algorithms were tested on a total of 11394 images from 4 other patients. For the SVM, 9 features were defined and extracted from each image. The CNN used the whole image as an input. Classification was based on 5 different outputs. The SVM achieved a sensitivity of 0.88 and a precision of 0.65 for DRs on the test data. For the CNN this was 0.96 and 0.42, respectively. Both models seem to have overfit on the underrepresented classes. Finally, the models were applied to data of 4 additional patients for comparison with human observers. For both models, the agreement with human observers was comparable to the inter-rater agreement for the spike and ripple frequency bands. We conclude that both models can be applied for a more efficient analysis of SPES.

At the second part of this research, we investigated the possibility of a CNN to find features that can distinguish between pathological and physiological DRs in TF-SPES images. The model was trained on 662 images and tested on 74 images, gathered from 8 different patients. All images contained DRs and were labeled as originating from either inside or outside the seizure onset zone (SOZ). The model achieved a sensitivity of 0.63 and a precision of 0.29 for DRs originating from the SOZ. These unsatisfactory results can be due to the low amount of data. Alternatively, it is suggested that the difference in pathological and physiological DRs cannot be found in TF-SPES images.

List of abbreviations

CNN: convolutional neural network
DL: deep learning
DR: delayed response (to SPES)
ECoG: electrocorticography
EEG: electroencephalography
ER: early response (to SPES)
ERSP: event-related spectral perturbation
ESM: electrical stimulation mapping
EZ: epileptogenic zone
F: fast ripple
fMRI: functional magnetic resonance imaging
HFO: high frequency oscillation
ID: interictal discharge
iEEG: intracranial EEG
IEMU: intensive epilepsy monitoring unit
IZ: irritative zone
MEG: magnetoencephalography
ML: machine learning
R: ripple
ReLU: rectified linear unit
ROI: region of interest
S: spike
SEEG: stereo-EEG
SOZ: seizure onset zone
SPECT: single-photon emission computed tomography
SPES: single pulse electrical stimulation
SVM: support vector machine
TF-SPES: time-frequency analysis of SPES
X: no DR

Contents

General introduction

1.1 Epilepsy surgery

Epilepsy is one of the most common neurologic conditions with a prevalence ranging from 0.3 to 1.2% in developed countries [1]. Out of this population, 20-30% continue to have seizures despite treatment with antiepileptic drugs [2]. For patients with focal drug-resistant epilepsy, surgery can be considered. The goal of epilepsy surgery is to remove the epileptogenic cortex from the brain, thus producing seizure freedom. Studies show that 40-50% of the patients who underwent epilepsy surgery remain seizure free 10 years after surgery [3, 4].

To achieve a good surgical outcome, it is of crucial importance to delineate the epileptogenic zone (EZ) and map the functionality of the surrounding cortex prior to surgery. It should be noted that the concept of an epileptogenic network, rather than a zone, might be better able to describe the complexity of seizure dynamics [5]. However, it remains uncertain how this concept should be applied in the clinical practice of epilepsy surgery. Therefore, the practical definition of EZ as "the minimum amount of cortex that must be resected (inactivated or completely disconnected) to produce seizure freedom" is most often used in contemporary medicine [6].

Lüders et al. define five different zones that can be directly measured and that are used as an indication of the location and extent of the EZ. These zones are the irritative zone (IZ), seizure-onset zone (SOZ), symptomatogenic zone, epileptogenic lesion and functional deficit zone. In their view, the IZ, which is the area of cortex which generates interictal discharges (IDs), is usually more extensive than the EZ, whereas the SOZ, the area of cortex which initiates clinical seizures, is a subset of the EZ [6].

Magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), single-photon emission computed tomography (SPECT) and electroencephalography (EEG) are noninvasive technologies that can assist in the presurgical evaluation [6]. If the noninvasive information is inconclusive or divergent, intracranial EEG (iEEG) is often done to measure activity and connectivity of the brain with electrodes directly on or in the neocortex [7]. iEEG can be done intra-operatively, with the placement of a subdural grid of electrodes, sometimes in combination with depth electrodes. This method is called electrocorticography (ECoG). The placement of invasive electrodes also provides an extra opportunity for determining functionality in the underlying cortex with electrical stimulation mapping (ESM). In

this method, part of the cortex is temporarily disconnected by high frequency electrical stimulation. Meanwhile, tests are performed on the patient to see if there is any loss of function, for example in language.

In a situation where there is need to intracranially capture seizure onset or when ESM is not feasible during surgery, a prolonged ECoG recording can be done through an independent implantation multiple days prior to the potential resection. This extra-operative iEEG creates extra time for extensive testing and capturing spontaneous ictal and interictal activity. It does, however, require an extra surgical procedure, with its own risks and morbidities [7]. An alternative extra-operative iEEG method, called stereo-EEG (SEEG), uses intracranial depth electrodes, which can be inserted through small burr holes, thus not requiring a craniotomy. Although the spatial sampling in SEEG is limited compared to ECoG, it does provide an opportunity for bilateral sampling of multiple deep structures.

1.2 SPES

In 2002, Valentín et al. introduced a method to provoke a response from the cortex by stimulation of intracranial electrodes, called single pulse electrical stimulation (SPES) [8]. The stimulations consist of a single block pulse with a duration of 1 ms, which is repeated every 5 seconds. The goal is to identify the limits of epileptogenic cortex. Here, epileptogenic cortex is defined as "the cortex that has the capacity to originate spontaneous seizures, presumably because it is abnormally hyperexcitable" [8]. This definition is based on the hypothesis that epilepsy arises from a cortical imbalance between excitatory and inhibitory mechanisms, meaning that epileptogenic tissue would have a more extreme response to SPES compared to healthy brain tissue. It should be noted that in this definition, epileptogenic cortex is very similar to the SOZ plus, what Lüders et al. call, the *potential* SOZ, which together form the EZ [6].

Two different kind of responses to SPES were observed, namely early responses (ERs, <100 ms after stimulation) and delayed responses (DRs, >100 ms after stimulation). See Figure 1.1 for an example. Valentín et al. designated the ERs to be a normal response of human cortex to stimulation and the DRs to be a pathological phenomenon. Both the areas where DRs are recorded as those which give rise to DRs when stimulated are called "abnormal SPES areas" and are related to the SOZ [8, 9]. In the past decade, their group has repeatedly shown that SPES is of added value in presurgical evaluation of epilepsy patients, especially when no on-going spontaneous (inter)ictal activity is present [8, 9, 10, 11, 12].

1.3 HFOs and SPES

High frequency oscillations (HFOs) recently emerged as a new biomarker for epileptogenic tissue [13, 14]. They consist of oscillations in the EEG above 80 Hz and are subdivided in ripples (80-250 Hz) and fast ripples (250-600 Hz). It is suggested that HFOs are a more reliable marker of the SOZ than spikes [13]. Also, it has been shown that HFOs can be reliably evoked by electrical stimulation [15]. This would save the time of waiting for spontaneous

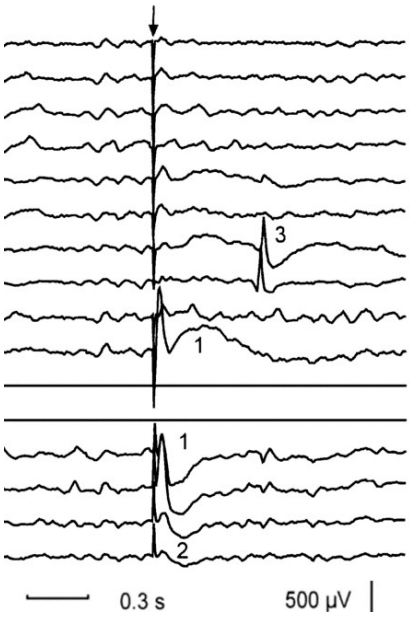


Figure 1.1: Example of SPES. Shown is 2 seconds of iEEG. The arrow at the top indicates the stimulation artifact. The two flat lines are the stimulated electrodes. 1 and 2 show ERs, right next to and further away from stimulation site. 3 shows a DR. (Adapted from [8]).

HFOs to occur. SPES provides an opportunity of evoking HFOs. However, with a visual analysis in the time domain, the components above the spike frequency range are easily missed. Therefore, Van 't Klooster et al. proposed an alternative method for the analysis of SPES, which includes the construction of an event-related spectral perturbation (ERSP) image around the stimulation [16]. This time-frequency analysis of SPES (TF-SPES) allows the observer to quickly see in what frequency bands the DRs contain activity (see Figure 1.2).

Similar to interictal spikes, the observed DRs in SPES are not 100% specific for the EZ. For example, in [9], 5 of the 7 patients in whom only part of the abnormal SPES areas was removed, nonetheless had a favorable outcome. The inclusion of HFOs in the analysis seems to increase this specificity. Van 't Klooster et al. found a 79% specificity of fast ripples for the SOZ, which was 17% for spikes [16]. However, compared to spikes, fast ripples showed a lower sensitivity (100% vs 67%). Moreover, HFOs also occur as a physiological phenomenon and currently SPES is not able to differentiate between pathological and physiological HFOs [13, 17]

1.4 SPES analysis and machine learning

In a standard SPES protocol in the UMC Utrecht, 10 pulses of 4 or 8 mA with a duration of 1 ms are given over all successive electrode pairs. This is shown in Figure 2.1. For each stimulation, one ERSP image is computed for every electrode. This can easily result in over

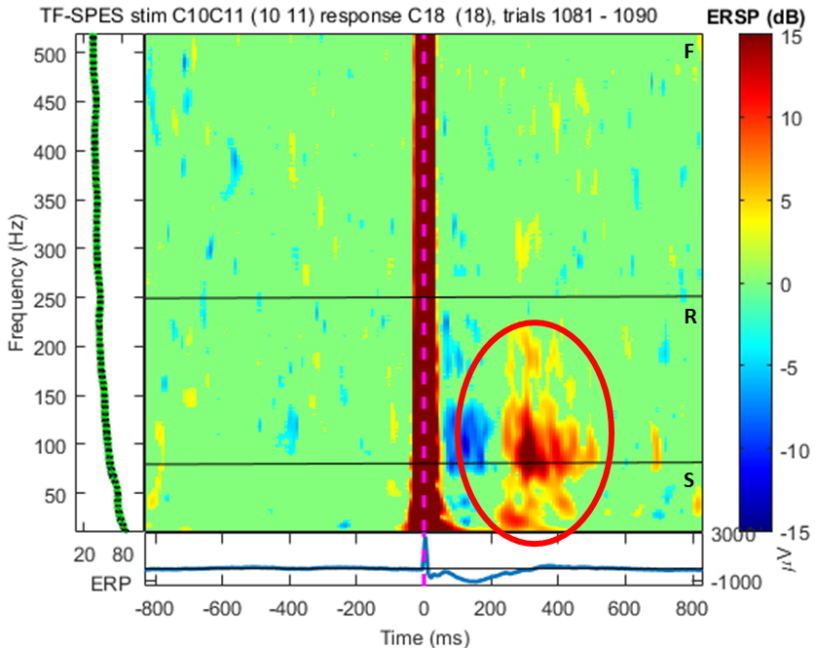


Figure 1.2: Example of an TF-SPES image with a DR. At 0 ms the stimulation artifact is visible. The DR is shown in the red ellipse. The black horizontal lines divide the ERSP into three different frequency bands, namely spike (S), ripple (R) and fast ripple (F). The plot left of the ERSP is the mean spectrum and the plot below the ERSP is the event-related perturbation in time-domain.

3000 images per patient which makes the visual analysis of TF-SPES very time consuming.

Previous attempts for automated classification of SPES responses were unsatisfactory (unpublished work). These attempts aimed to mimic human pattern recognition based on certain quantitative features extracted from the ERSP image, such as the maximum and cumulative of the increased power, and features based on ridge detection of the image. The main reason that these attempts failed was that the chosen features turned out to be unable to distinguish evoked responses from (muscle) artifacts that are fairly obvious to the human eye. Apparently, it is very hard to find features that mimic the human pattern recognition. Moreover, manually defining a threshold for these features makes it even harder to create an automated detection system. This is where machine learning (ML) might help. In ML, 'learning' means to improve at a certain task, as measured by a certain performance measure, by use of given examples [18]. One application of ML is to program a computer to construct a model, or classifier, for separating data of different categories [19]. In supervised ML, each given example consists of a set of features and its category, or class. The computer takes the features of the given examples as inputs and optimizes the parameters of the model in such a way that the outputs are best fitted to the classes belonging to the examples. Ideally, this optimization is done in such a way that the model is able to generalize over new examples that were not used for training.

The support vector machine (SVM) is a very commonly used ML classifier described by

Vapnik in 1995 [19]. An SVM separates the different classes by constructing a hyperplane in feature space that has a maximal margin between the so-called support vectors, which are the samples closest to the hyperplane.

Although ML makes it easier to combine multiple features, the quality of the automated classification still depends on the discriminative power of the given features. As mentioned previously, the crafting of features that mimic human pattern recognition is not an easy task. A recent development in ML, called deep learning (DL), might be a solution to this problem. In DL, there are multiple 'hidden' layers between the input and output of the model which transform the input to an increasingly more abstract representation, followed by classification at the final part of the model [20]. Thus, the extraction of patterns that are useful for classification is part of the learning and the machine can be fed with raw inputs. This is similar to human pattern recognition, where our brain unconsciously extracts these useful patterns out of an image. In fact, the convolutional layers of a convolutional neural network (CNN), which is a form of DL that is commonly used for image classification [21, 22, 23, 24], are directly based on the functional architecture found in a cat's visual cortex [20]. The fact that there are multiple layers between the input and output layer is why it is called 'deep' learning, as opposed to 'shallow' classifiers, such as the SVM.

The definite breakthrough of the CNN was the automated handwritten character recognition system of LeCun et al. [21]. They showed that, for character recognition, carefully designed learning machines that operate directly on pixel images can do a better job than discrimination based on hand-crafted features. Also in medical imaging, pixel-based ML is increasingly more popular [22]. Furthermore, as a DL network extracts its own features from the input data, it may find patterns that are useful in a classification task for which the human eye is not trained. A good example of this is given by Van Putten et al., who trained a CNN to predict the patient's sex solely based on EEG recordings [25]. Beforehand, it was unknown that there are features present in the EEG that are discriminative for the patient's sex. Likewise, DL may succeed in discriminating pathological DRs from physiological ones, based on features currently unknown by the human observer.

A downside, however, is that this type of classification has a relatively long training time due to the high dimensionality of the input data. Furthermore, besides a large training set, also considerable computational power and a good learning machine architecture is needed for this ML method to perform as desired.

1.5 Research question and objective

Based on the previous, we can broadly state two research objectives:

- To significantly reduce the amount of time and effort needed for the analysis of SPES, using machine learning
- To discover the possibilities of machine learning in distinguishing between pathological evoked responses and physiological ones

With these objectives as a starting point, we define four research questions:

⇒ Can machine learning detect DRs with high sensitivity and precision?

⇒ Can machine learning adequately distinguish DRs with different frequency components?

⇒ Can deep learning outperform a shallow classifier in SPES responses classification?

⇒ Can deep learning distinguish between DRs originating from inside and outside SOZ?

ML for the classification of DRs

2.1 Objective

The first goal was to use ML to automatically detect and classify DRs in the TF-SPES images. We trained both an SVM and a CNN for this purpose. Data of 19 patients were selected for training and testing the models. The models were applied to data of 4 additional patients for a comparison of the agreements with human observers and the inter-rater agreements. For the mathematical details of the ML models, we refer to the Appendix.

2.2 Methods

2.2.1 SPES data acquisition and processing

All selected patients clinically underwent SPES in the past four years during their stay in the intensive epilepsy monitoring unit (IEMU) of the UMC Utrecht. These patients exclusively had chronic ECoG, meaning that no SEEG patients were included in this study. Their iEEG was recorded with MicroMed LTM 128/64 express EEG headbox with integrated programmable stimulator (MicroMed, Mogliano Veneto, Italy). The sample frequency was 2048 Hz. SPES was performed with 10 block pulses of 8 mA and a duration of 1 ms that were supplied to all successive electrode pairs (see Figure 2.1). When stimulating in the (assumed)

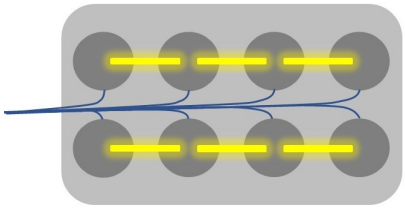


Figure 2.1: Stimulation representation. Shown is a 2x4 grid, where the dark gray circles represent the electrodes. The yellow bars represent the given stimulation locations in a standard SPES protocol.

motor or sensory cortex, the amplitude was lowered to 4 mA. The data was further processed with Matlab (Matlab R2016a, Mathworks Inc. MA, USA) using the EEGLAB toolbox [26]. The power spectrum $P(f,t)$ was calculated in the following way:

$$P(f,t) = \frac{1}{n}\sum_{k=1}^{n} \mid F_k(f,t) \mid^2 \tag{2.1}$$

where $F_k(f,t)$ is a sinusoidal wavelet transform of trial k at frequency f and time t, beginning with a 3-cycle wavelet and gradually increasing to a factor of 0.8 at the highest frequency. n is the number of stimulations given at each stimulation pair and was equal to 10. The frequency ranged from 10 to 520 Hz and the time range was taken from $t = -1$ s to $t = 1$ s where $t = 0$ is the time of stimulation. This resulted in 511x200 sized matrices, with a frequency resolution of 1 Hz/sample and a time resolution of approximately 8.3 ms/sample. Hereafter, the baseline power was calculated for each frequency, i.e. the mean of each spectral estimate from $t = -1000$ ms to $t = -200$ ms (or $t = -100$ ms in 4 patients). This baseline power was subtracted from the whole power spectrum. Each value was set to 0 if it did not significantly deviate from the baseline, based on a bootstrap method with 200 bootstrap replications and $\alpha = 0.05$. This resulted in the final ERSP in which the values are given in dB. The images were constructed with a *jet* colormap ranging from -15 to $+15$ dB. Figure 1.2 shows an example.

Scoring of images

All TF-SPES images of the selected patients were scored by two human observers. They scored the images on presence of DRs in three different frequency bands, namely spikes (S, 10-80 Hz), ripples (R, 80-250 Hz) and fast ripples (F, 250-520 Hz). For example, if a DR shows activity in all three frequency bands, the label would be SRF (spike, ripple and fast-ripple). If there is activity only in the ripple band, the label would be R. As the different frequency bands are not mutually exclusive, there are $2^3 = 8$ different kind of scoring possibilities for an image. The example shown in Figure 1.2 was scored as SR.

Currently, two experts score each patient independently, after which the level of agreement is assessed within each frequency band with a Cohen's kappa. If $\kappa < 0.4$, the scores of that specific frequency band are discarded. The same threshold was used in this research for the final comparison of ML models vs human observers. However, for the selection of our train and test data, we ignored the inter-rater agreement of the images.

2.2.2 Training and testing of models

Data selection

In the images of the 19 patients used for training and testing purposes, the visually set labels were used as a ground truth. We excluded the images upon which the observers had disagreed. This means that we discarded an image if, for example, observer 1 scored S and observer 2 scored SR, even though they agreed upon the presence of a DR in the spike frequency band.

We also excluded images that were labeled as either F, SF or RF, because these were very rare in the remaining data. This left us with five different labels, namely X (no DRs), S, R, SR and SRF. Instead of the color value of the scored images, the dB values were used as a starting point. Also, the 200 ms around stimulation was removed, resulting in 511x178 sized matrices (Figure 2.2 a). This was done to exclude the stimulation artifact, which does not contain any useful information, and the ERs, which are irrelevant for scoring the DRs.

The data was divided into a training set of 15 patients and test set of the remaining 4 patients. The test set was chosen in such a way that the representation of each class was similar to that of the training set. We chose to have separate patients in the test set, rather than to do stratified sampling on the classes, to see if the models would be generalizable to patients it has never seen before. The distribution of classes in the test and training set are shown in Table 2.1.

Table 2.1: data samples in training and test data

	X	S	R	SR	SRF	Total
Train	43,653	1,191	301	1,873	179	47,197
Test	10,428	413	78	423	52	11,394

X: no DR. S: spike. R: ripple. SR: spike-ripple. SRF: spike-ripple-fast ripple.

Class-weights

The images containing no DRs were overrepresented in the data. To deal with this skewness, we used a balanced class-weight for both algorithms. This means that during training, the loss for a certain class is multiplied with a weight that is inversely proportional to the fraction of that class in the total data. Thus, a mistake in the underrepresented classes is more heavily punished than a mistake in the overrepresented class.

Performance measure for optimization

We performed cross-validation for optimization of the hyper-parameters of the models. The so-called macro F1-score was used as the metric for cross-validation:

$$F1 = 2 * \frac{P * S}{P + S} \tag{2.2}$$

where P is precision and S is sensitivity:

$$P = \frac{TP}{TP + FP}, \quad S = \frac{TP}{TP + FN} \tag{2.3}$$

with TP being true positives, FP false positives and FN false negatives. The macro F1-score is the unweighted average of the F1-scores calculated separately for each class. This means that the F1-score of an underrepresented class is equally important as the score of an overrepresented class in optimization of the hyper-parameters. Thus, the macro F1-score helps dealing with the unbalanced classes.

Evaluation and comparison of models

To evaluate the detection of DRs, we calculated the sensitivity and precision of both models for all DR classes (S, R, SR and SRF) taken together. For the distinction between different kind of DRs, we also calculated the sensitivity and precision for all classes separately.

For statistical comparison of both models we separated the number of samples in the test set that both models classified correctly, that both models classified wrongly or that were correctly classified by just one model, as shown in Table 2.2. This was also done for only the DR classes in the test set.

Table 2.2: Contingency table

		CNN	
		Correct	Wrong
SVM	Correct	N_{++}	N_{+-}
	Wrong	N_{-+}	N_{--}

To test the hypothesis H_0 that both models score equally well, we applied McNemar's test on both contingency tables [27]:

$$\chi^2 = \frac{(|N_{+-} - N_{-+}| - 1)^2}{N_{+-} + N_{-+}}, \tag{2.4}$$

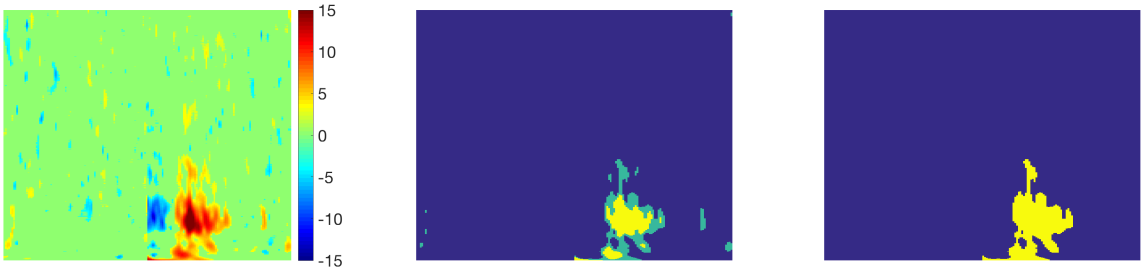
where the -1 is added to correct for continuity. H_0 is rejected when $p < 0.05$.

2.2.3 SVM for image classification

The image processing and training of SVM was done in Matlab (Matlab R2016a, Mathworks Inc. MA, USA).

Image segmentation and definition of features

First, a segmentation technique called hysteresis thresholding was used to select regions of interest (ROIs) in the images. A way to look at this technique is that it starts from regions that are above a certain threshold (T_h), and from there spreads out until it reaches a second, lower threshold (T_l). See Figure 2.2 for an example. From the selected ROIs we extracted nine different features, which were ought to represent the ROIs on aspects on which human observers base their decision in scoring the images. The features are listed in Table 2.3. When no ROI was selected, the values of the features were set to 0. We divided each image into baseline and response, and extracted the features from both parts separately. We decided to ignore the ROIs that had their highest frequency < 30 Hz and timing > -130 ms or < 130 ms, as these ROIs commonly turned out to be attached to the stimulus artifact. Hereafter, the features were standardized to have zero mean and standard deviation 1.



(a) (b) (c)

Figure 2.2: Example of hysteresis thresholding. (a): input matrix. (b): Thresholding applied to the input matrix, with $T_h = 7$ and $T_l = 4$. Blue$< T_l$, yellow$\geq T_h$ and $T_l \leq$green$< T_h$. (c): Resulting ROI shown in yellow.

Table 2.3: Extracted features for SVM classifier

	Baseline	Response
Total Power	x	x
Highest frequency	x	x
Lowest frequency	x	x
Duration	x	x
Timing		x

Total Power: the sum of all values in the ROIs. Highest frequency: the highest frequency present in any of the ROIs. Lowest frequency: The lowest frequency present in any of the ROIs. Duration: The median with of all ROIs. Timing: The x-coordinate of the centroid of the latest ROI.

Model choices

We used a multi-class SVM with Gaussian radial basis function kernel. The output coding was set to one-versus-one, which generally seems to perform better than one-versus-all [28]. This means that for each possible pair of classes, a binary classifier is trained, i.e. $\frac{1}{2}(K-1)K$ binary classifiers, where K is the number of classes. In applying the model, an observation is assigned to that class for which the sum of losses of all binary classifiers is smallest [28].

Optimization of hyper-parameters

We defined three so-called model hyper-parameters, namely T_h, T_l and the C parameter of the SVM. C regulates the number of support vectors used for constructing the hyperplane: higher C means less support vectors. Generally, more support vectors means that the hyperplane is less strictly fitted to the data, which makes the classifier more generalizable. The hyper-parameters were optimized using grid search with 10-fold cross-validation on the train set. We used the following range of values for C: 2^α with $\alpha \in \{-4, -2, 0, 2, 4, 6, 8\}$. The ranges

of thresholds for segmentation were: $T_l \in \{2, 3, 4, 5, 6\}$ and $T_h \in \{7, 8, 9, 10, 11\}$.

Training of final model

After the cross-validation, a final model was trained on all training data, using the optimal hyper-parameters. This model was applied to the test data for evaluation of its performance.

2.2.4 CNN for image classification

We used Python (Python Software Foundation, version 3.5) for constructing the CNN with Keras library and TensorFlow backend [29]. Training was done on two CUDA-enabled NVIDIA Tesla P100 GPUs.

Model choices

The 511x178 sized matrices were used as inputs for the model. The labels (X, S, R, SR and SRF) were set to a one-hot representation. The complete structure of the CNN is visualized in Figure 2.3. The convolutional layers 'scan' its inputs with a certain window, called the kernel, where every step in this 'scan' has the same weights. The max-pooling layers are subsampling layers that scale the feature maps down by taking the maximum value of a certain window size. Not depicted in the figure are the dropout layers that were after both max-pooling layers. Dropout layers prevent the network from overfitting and are used instead of a regularization term in the objective function. The activation function of the output layer was set to softmax, whereas the fully connected layer and both convolutional layers consisted of rectified linear units (ReLU). Softmax transforms the outputs to a categorical probability distribution, so that each output node has a value between 0 and 1, and all output nodes sum up to 1

Cross-entropy was chosen as the loss function. The loss function is the performance measure of the model. During training, the model improves by minimizing the loss that is the output of the loss function. The number of samples in each learning batch was set to 250. ADADELTA was used for setting the learning rate parameter. ADADELTA automatically initializes and updates the learning rate η over time [30]. In applying the model, an observation is assigned to that class for which output vector $\mathbf{y}$ has the largest value.

Optimization of hyper-parameters

The number of nodes in the two convolutional layers ([2, 3, 4, 5]) and the dropout rate of the dropout layers ([0.1, 0.3, 0.5]) were optimized with grids search on the train set. This was done by 1-fold cross-validation, based on a cross-validation set that consisted of 20% of the train set and was constructed with stratified sampling. For optimization, 10 epochs were used for training.

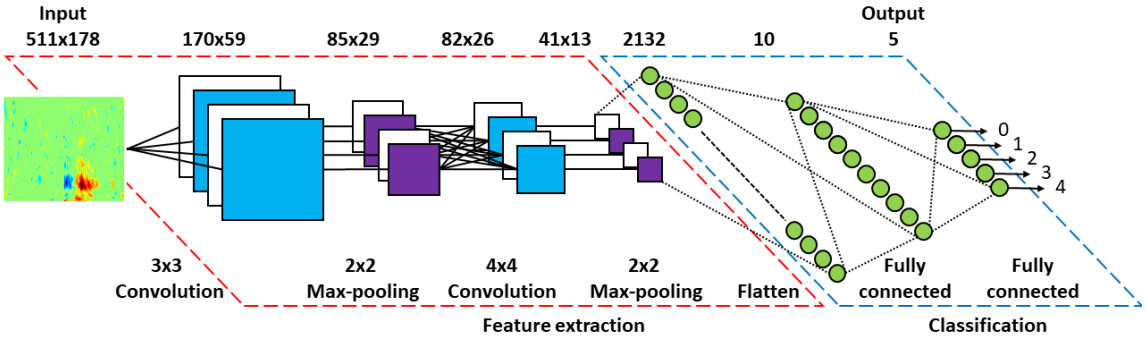


Figure 2.3: Network structure. On the bottom are descriptions of the different layers with the used kernels. The first convolutional layer had a stride of 3, other layers had 0 stride. Not depicted are the dropout layers that followed both pooling layers. Shown above the feature extraction compartment are the sizes of the feature maps that are the outputs of each layer and inputs for the next layer. Above the classification section are the number of nodes.

Training of final model

Finally, the algorithm was trained on the whole training set with optimal hyper-parameters and tested on the test set. Here, we used 40 epochs for training.

2.2.5 Comparison with human observers

Finally, we applied both models to all TF-SPES images of four additional patients, not used for training or testing purposes. Cohen's kappa was calculated within each frequency band, as is currently done in clinical practice. We compared the means of kappas between each model and both human observers with the inter-rater agreements, to see whether the models could function as an observer in practice. Here, we regarded $\kappa \geq 0.4$ as an acceptable agreement.

2.3 Results

2.3.1 Results of SVM

Optimization of hyper-parameters

The optimal hyper-parameters, resulting from the 10-fold cross-validation, were: $T_l = 5$, $T_h = 8$ and $C = 2^2 = 4$.

Detection of DRs

On the test set, sensitivity and precision for the DRs taken all together was 0.88 and 0.65, respectively. For the train set, this was 0.98 and 0.68.

Distinction between different kind of DRs

Table 2.4 and 2.5 show the confusion matrices of the SVM applied to the train and test data, respectively. Figure 2.4 shows the sensitivity and precision of each class individually.

Table 2.4: Confusion matrix SVM on train set

		Predicted					
		X	S	R	SR	SRF	Total
True	X	42025	861	349	397	21	43653
	S	50	1104	0	35	2	1191
	R	15	2	277	7	0	301
	SR	19	90	89	1624	51	1873
	SRF	0	4	0	9	166	179
	Total	42109	2061	715	2072	240	47197

The correctly predicted samples are on the diagonal, shown in gray. X: no DR. S: spike. R: ripple. SR: spike-ripple. SRF: spike-ripple-fast ripple.

Table 2.5: Confusion matrix SVM on test set

		Predicted					
		X	S	R	SR	SRF	Total
True	X	9964	296	66	101	1	10428
	S	57	342	0	14	0	413
	R	14	2	37	25	0	78
	SR	31	60	20	291	21	423
	SRF	13	3	0	18	18	52
	Total	10079	703	123	449	40	11394

The correctly predicted samples are on the diagonal, shown in gray. X: no DR. S: spike. R: ripple. SR: spike-ripple. SRF: spike-ripple-fast ripple.

2.3.2 Results of CNN

Optimization of hyper-parameters

The optimal number of nodes for both convolutional layers of the CNN was 4. This can also be seen in Figure 2.3. The resulting network had a total of 21,685 trainable parameters. The two dropout layers had an optimal dropout rate of $p = 0.1$ and $p = 0.3$ respectively.

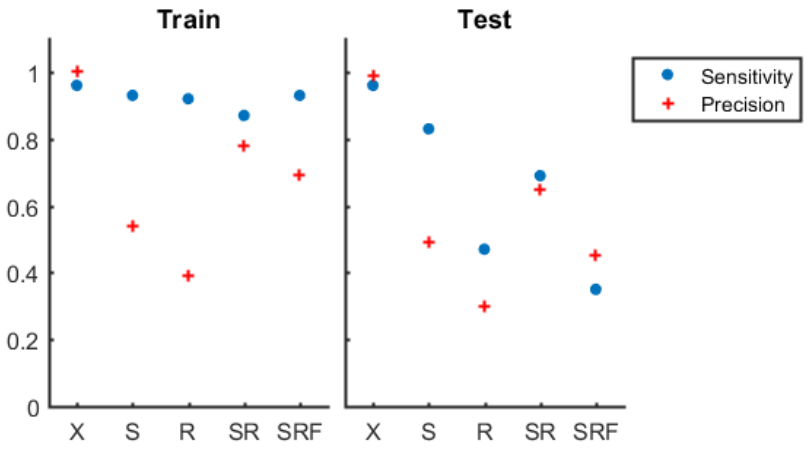


Figure 2.4: Sensitivity and precision SVM on test and train set. X: no DR. S: spike. R: ripple. SR: spike-ripple. SRF: spike-ripple-fast ripple.

Detection of DRs

In the test set, sensitivity and precision for the DRs taken all together was 0.96 and 0.42, respectively. For the train set, this was 0.98 and 0.40.

Distinction between different kind of DRs

Figure 2.5 shows the learning curve of the training and test set of the final trained CNN. Table 2.7 shows the confusion matrix of the CNN applied to the test data and Figure 2.6 shows the sensitivity and precision of each class individually.

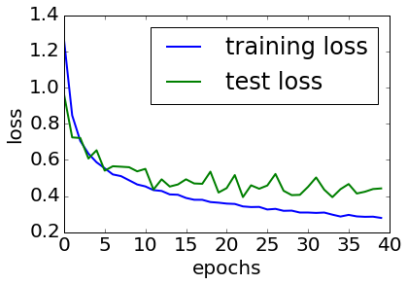


Figure 2.5: Learning curve of the CNN. Shown is the loss after each epoch of training for both train and test set (see Equation 5 in Appendix A). The test loss no longer seems to decline after 40 epochs, meaning that the model has stopped improving on the test set by training on the train set.

Table 2.6: Confusion matrix CNN on train set

		Predicted					
		X	S	R	SR	SRF	Total
True	X	38463	2837	1479	799	75	43653
	S	39	1133	0	19	0	1191
	R	1	0	300	0	0	301
	SR	26	74	76	1682	15	1873
	SRF	0	0	0	0	179	179
	Total	38529	4044	1855	2500	269	47197

The correctly predicted samples are on the diagonal, shown in gray. X: no DR. S: spike. R: ripple. SR: spike-ripple. SRF: spike-ripple-fast ripple.

Table 2.7: Confusion matrix CNN on test set

		Predicted					
		X	S	R	SR	SRF	Total
True	X	9130	798	337	145	18	10428
	S	33	359	0	21	0	413
	R	4	2	60	12	0	78
	SR	4	53	15	344	7	423
	SRF	0	9	1	21	21	52
	Total	9171	1221	413	543	46	11394

The correctly predicted samples are on the diagonal, shown in gray. X: no DR. S: spike. R: ripple. SR: spike-ripple. SRF: spike-ripple-fast ripple.

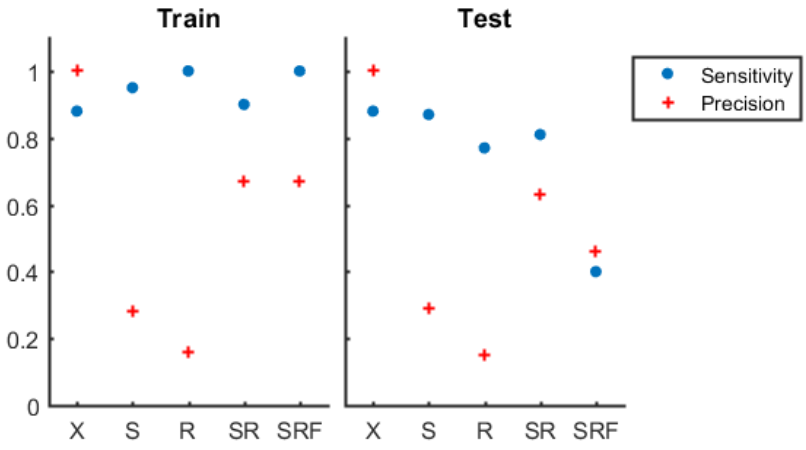


Figure 2.6: Sensitivity and precision CNN on test and train set. X: no DR. S: spike. R: ripple. SR: spike-ripple. SRF: spike-ripple-fast ripple.

Feature extraction

To see whether the features that the model learned to extract resemble the features that were defined for the SVM, the activations of some examples were reviewed. Figure 2.7 shows an example of the activations of the network to an input image.

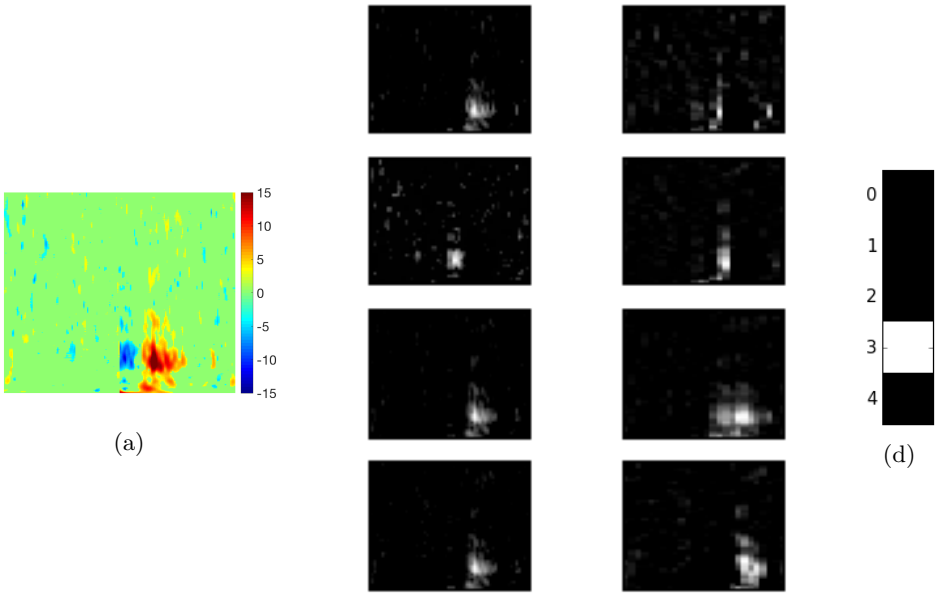


Figure 2.7: Example of activations of CNN. (a): input matrix. (b): Feature maps of first convolutional layer. (c): Feature maps of second convolutional layer. (d): Softmax output. Both (b) and (c) are grayscale images scaled on its min and max values. These layers represent the features extraction of the model, meaning that they show what the model deems important in the input image for classification. In (d), 0 to 4 represent the output labels and the grayscale ranges from 0 to 100%. In this example, the network is 100% sure the label is 3, i.e. SR. The input image was part of the train set.

2.3.3 Comparison of models

Table 2.8 shows the contingency tables for all test data and for only the DRs in the test data.

Table 2.8: Contingency tables

		CNN	
		Correct	Wrong
SVM	Correct	9552	1100
	Wrong	362	380

		CNN	
		Correct	Wrong
SVM	Correct	604	84
	Wrong	180	98

Left: all data in test set. Right: only DR classes in test set.

McNemar's test, applied to the contingency tables, was as follows for all test data:

$$\chi^2 = \frac{(|1100 - 362| - 1)^2}{1100 + 362} = 371.52. \tag{2.5}$$

And for only the DRs:

$$\chi^2 = \frac{(|84 - 180| - 1)^2}{84 + 180} = 34.19. \tag{2.6}$$

Both give $p < 0.001$. This shows that, reviewing Table 2.8, the SVM is overall significantly more correct than the CNN, but the CNN is significantly better in scoring the different kind of DRs.

2.3.4 Comparison with human observers

Table 2.9 shows Cohen's kappa per frequency band for the four additional patients. Displayed is the kappa between both human observers and the mean kappa of model vs. human observer 1 and model vs. human observer 2. In red are $\kappa < 0.4$.

Table 2.9: Kappas

		Inter-rater	Mean SVM vs. human	Mean CNN vs. human
S	Pat 1	0.43	0.46	0.33
	Pat 2	0.50	0.44	0.52
	Pat 3	0.75	0.72	0.70
	Pat 4	0.58	0.50	0.50
R	Pat 1	0.37	0.34	0.19
	Pat 2	0.50	0.43	0.55
	Pat 3	0.76	0.73	0.70
	Pat 4	0.52	0.52	0.46
F	Pat 1	0.11	0.11	0.05
	Pat 2	0.81	0.13	0.50
	Pat 3	0.64	0.32	0.54
	Pat 4	0.47	0.33	0.20

S: spike. R: ripple. F: fast ripple. Pat: Patient. Numbers in red are $\kappa < 0.4$.

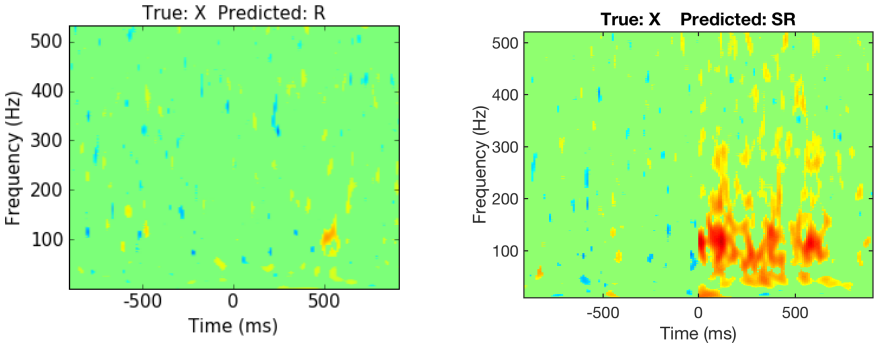
2.4 Discussion

We constructed two ML models with high sensitivity (0.88 and 0.96 on test set) for the detection of DRs in TF-SPES images. Both models achieved kappas with human observers that were comparable to the inter-rater agreements in the spike and ripple range. Although the models are not ready to fully replace the human observer, they can already be applied to assist in the analysis of SPES.

2.4.1 Performance of models in scoring DRs

Detection of DRs

Both models show a much higher sensitivity than precision for DRs. This is the case for both the train as the test set, showing that the models did not overfit on this aspect. It seems that this effect is stronger for the CNN than for the SVM. These differences in sensitivity and precision for DRs are very likely due to the balanced class-weights. The class-weight for X, being abundant in the train set, is very small (Equation 1 in Appendix A), which means that a mistake in classifying this class is punished very lightly compared to a mistake in the more rare DR classes. Figure 2.8 shows an example of how the CNN overestimates the response activity and how the SVM misinterprets artifact activity in scoring an image of the test set. On the other hand, one could argue that the models, especially the CNN, are not necessarily wrong in these examples. The limitation of the used ground truth is discussed further on.



(a) (b)

Figure 2.8: Example of misclassification of CNN (a) and SVM (b) in test set, based on the consensus of human observers.

In Figures 2.4 and 2.6, the imbalance between sensitivity and precision can be observed for the separate DR classes. The confusion matrices (Tables 2.4, 2.5, 2.6 and 2.7), show that the relatively low precisions are indeed mainly due to the images with label X that are scored by the models as one of the DR classes. However, for both models there exists a relatively large difference in sensitivity for R and SRF between train and test set. It is possible that the patients in the test set show DRs that look very different from those in the train set, which makes it hard for the models to generalize over patients, but more likely this is due to the class-weight as well. The classes R and SRF together formed only 1.02% of the total training data, so a mistake on one of these DRs receives a hard punishment. It is likely that this resulted in the models overfitting on these specific classes, which is why they seemed to have problems generalizing on the test set. To find a better balance between sensitivity and precision for DRs, the class-weights could be optimized during cross-validation.

A different solution to the class imbalance, that might work better than setting class-weights, is to simply gather more data of the minority classes. This can be done by including more patients or by creating synthetic training data. One method to synthetically create more

input data is using label-preserving transformations [24]. For example, flipping the baseline of the TF-SPES images horizontally would create a new image with the same label. However, the added value of this new input image was considered to be marginally, as the network has already learned that the most important information is in the response part of the image, which stays unchanged. A different method for synthetically creating more input data is called synthetic minority over-sampling technique (SMOTE) [31]. In this method, new samples are created by projecting randomly on the lines between a sample and its k-nearest neighbors in feature space. This method could be considered for future research.

Distinction between different kind of DRs

When it comes to distinguishing between the different kind of DRs, Tables 2.5 and 2.7 show that both models relatively often score SRF as SR. Apparently, there are not enough SRF images in the training set to get the models to learn the border between the ripple and fast ripple frequency bands. It can be argued that it is 'unfair' that the human observers have foreknowledge on the borders between the different frequency bands whereas the models have to figure them out themselves. Also information on the response electrodes, known by the human observers during scoring of the images, was not taken into account in training the models. This included whether the electrode was part of the stimulation pair or whether it was considered to be a 'bad' channel due to a low signal-to-noise ratio. The images belonging to these electrodes should be discarded in future research. To overcome the problem of the 'unknown' borders between frequency bands, each band could be separated for both models. For the SVM this would mean that the features 'highest frequency' and 'lowest frequency' are discarded and the remaining features are extracted separately for the different frequency bands. For the CNN this would mean that each input matrix is split into three matrices, which are fed into three separate feature extraction compartments. These compartments would then be combined in the classification compartment of the network to form one output.

Comparison with human observers

It has to be noted that the consensus of two human observers is not a universal ground truth; a third human observer might disagree. Even though the human observers know the borders between the frequency bands, one still might argue whether the activity shown in, for example, the fast ripple band is 'intense' enough to be scored as present. Indeed, when reviewing some of the images, there is no clear answer to which observer is 'correct' (e.g. see Figure 2.9 (b)). For that reason, we applied both models to four additional patients, so we could compare the agreement within human observers with the agreements between human observer and model.

Similar to the inter-rater agreement, we considered $\kappa \geq 0.4$ as acceptable. The kappas can be reviewed in Table 2.9. It can be seen that when it comes to the spike and ripple frequency bands, both models perform very reasonably. Only once the CNN achieves a kappa below 0.4 when the inter-rater agreement is above 0.4. When reviewing some images, especially the CNN often seems to overestimate the response activity, similar to Figure 2.8 (a). Apparently, this is not dramatic for the kappas in the lower two frequency bands. However, when it comes

to the fast ripple band, especially the SVM seems to fail. To find a reason for this, we reviewed some images from Patient 2. For this patient, the difference in kappas in the F band was biggest. Some key examples are shown in Figure 2.9. Examples (a) and (b) show that the models have trouble in defining the border of the R and F frequency bands, as discussed previously. Although in (b), one might argue that the response has actually crossed that border, thus agreeing with both models. In example (c), the SVM seems to overestimate the baseline activity, which is much less pronounced than the response activity. The general problem of SPES with the presence of spontaneous interictal activity is discussed in section 4.1. In (d), the SVM seems to misinterpret some of the high frequency activity. This could be caused by the fact that the SVM has no way of 'knowing' whether the separate ROIs are at the same or at different timings. Also, the SVM cannot take 'light' activity, displayed in yellow and orange, into account, as it is below the hysteresis thresholds. It is probable that in this image, however, the 'light' activity does play a role for the human observers in scoring the image.

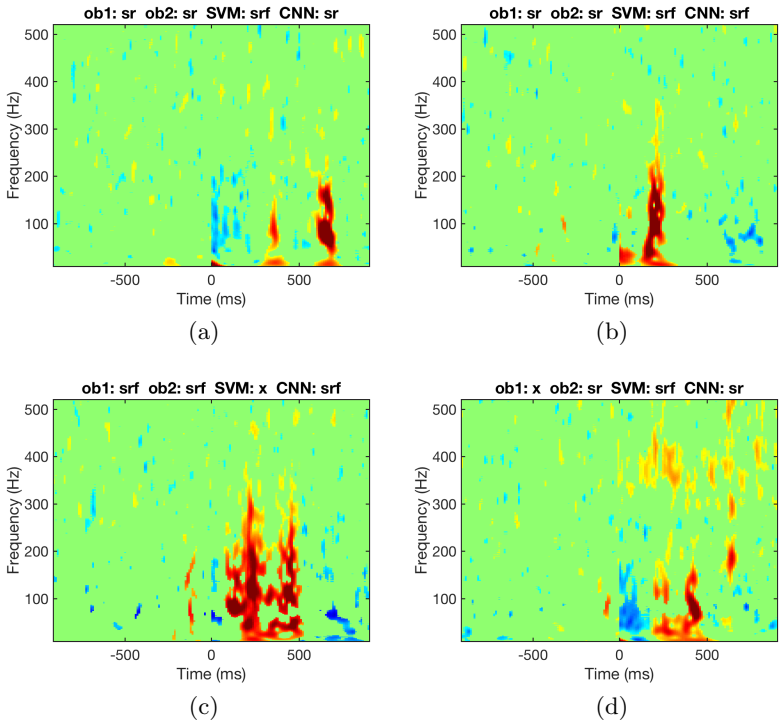


Figure 2.9: Examples of images from patient 2. The scores of the human observers (ob1 and ob2) and the two models are shown above the images.

2.4.2 SVM model considerations

Kernel function

We decided to use a Gaussian radial basis function kernel for the SVM, because we reasoned that for some features there is no linear way to distinguish between certain classes. For example, if the duration of a ROI is extremely short, it is likely to be an artifact. However, if the duration is too long, it also does not look like a genuine DR.

Choice of features

The features used for the SVM were chosen to represent aspects of the image on which the human observers base their decision. Obviously, five different types of features cannot entirely capture the complexity of the image, but we assumed that they represent the ROIs enough for scoring purposes. One problem that we encountered is that in the 'empty' images, where no ROI was selected, all features had value 0. As these images were abundant, the features had no normal distribution, which can cause the SVM to perform badly. An alternative was to a-priori score the empty images as X (no DR) and leave them out of the training set, but this still leaves you with images that have ROIs in response, but are empty in baseline. This may make it hard for the model to distinguish between baseline activity that is acceptable for a genuine DR (see Figure 2.9 (c)) and baseline activity that makes it hard to tell whether the activity after stimulation is spontaneous or evoked. In the latter case, the human observers score X. Subtracting the baseline features from the response features to form a new feature set should be considered.

2.4.3 CNN model considerations

Architecture and optimization

The structure of the network was roughly based on the architecture of LeNet-5 from [21]. We decided to only optimize the number of nodes and dropout rate with cross-validation, although it could be argued that also the number of convolutional layers and its kernel sizes could be optimized. However, the benefit of optimizing all hyper-parameters of the network was deemed marginally compared to the extra time required for cross-validation. Furthermore, the network was designed in such a way that the number of trainable parameters would not succeed the number of input images, to prevent overfitting. The addition of a third convolutional layer, followed by a dropout layer, could be considered. According to Lecun et al., the second convolutional layer typically detects motifs in an image, whereas the third convolutional layer may assemble these motifs into larger combinations [20].

Considering the longer training time usually needed for training a DL model, compared to a shallow classifier, we decided to use a 1-fold cross-validation instead of the 10-fold cross-validation used for the SVM. However, the GPUs used for training the model reduced the training time with such a large factor (approximately 48 times) that it could be considered

to increase the number of folds or range of hyper-parameters in future research. This would result in a better optimization, although it is expected that the performance would only improve marginally.

Figure 2.5 shows that the number of epochs for training the CNN was well chosen in the sense that test loss is not yet increasing, but seems to have stopped decreasing.

Classification

The labeling of images was based on the highest value of the output vector. Alternatively, additional constraints could be applied to the classification, such as a threshold for the highest value or for the difference between the highest and second highest value. By varying this threshold, a ROC curve could be set up of the performance of the CNN.

Feature extraction

Figure 2.7 shows an examples of the feature maps that are extracted from an input image. Remarkable is that the second node in the first convolutional layer seems to focus on the blue region of the input image, i.e. suppression of a certain frequency range after stimulation, whereas the human observers do not (consciously) take this into account for scoring. The activations in the other three nodes of the first convolutional layer rather look alike, which give the impression that less nodes also would have been sufficient. However, that conclusion can not be drawn from just one example. What the activations in the second convolutional layer represent and how they are weighted to eventually form the output is hard to interpret.

2.4.4 Conclusion

One objective of this research was to significantly reduce the amount of time and effort needed for the analysis of SPES. Two different ways to achieve this objective can be defined. The first is to assist the human observer by filtering out the vast majority of X images, thus leaving only a fraction of the total number of images for the human observer to score. A high sensitivity for DRs is needed for this purpose. In our opinion, the models, especially the CNN, performs well enough to fulfill this goal. A different approach to achieve this goal, is to only apply the hysteresis thresholding to the images and set the thresholds to the highest values that give 0 false negatives. In this way, one could easily discard all 'empty' images, which is the majority of all X images.

The second way to achieve the objective is to entirely replace the human observer in scoring the images. For this purpose, a high agreement with the human observers is needed. In reviewing the kappas, we can conclude that the models are not quite there yet, but especially the CNN seems to have the potential to achieve that goal.

Although the SVM seems to have scored better on the train and test set compared to the CNN, the aforementioned limitations of the SVM give us a preference towards the CNN. The CNN scored better on the kappa analysis and its main limitation seems to be the that it

too easily scores 'light' activity as a DR. It is expected that the suggested improvements, especially the creation of more input data containing DRs, will bring this limitation to a minimum.

DRs and the SOZ

3.1 Objective

The second objective was to investigate whether ML could distinguish between pathological and physiological SPES responses. For this purpose, we trained a CNN to classify DRs as originating from either inside or outside the SOZ. ECoG data of the same patients included in Chapter 2 were used.

3.2 Methods

3.2.1 Data selection

An expert neurologist reviewed the seizures of each patient for presence of gamma activity and to determine the SOZ electrodes. The reviewed seizures were captured during the patient's stay at the IEMU and recorded as described in section 2.2.1. Hereafter, we selected only the patients in whom the ECoG showed gamma activity at seizure onset. This was done to be more certain that in our data we actually had electrodes directly on top of the SOZ. The determined SOZ electrodes were used as the ground truth. We excluded the patients where the SOZ was hard to determine or who had a diffuse SOZ. This left us with 8 patients.

3.2.2 Statistical relation to SOZ

First, we computed the sensitivity and specificity for the SOZ of the electrodes showing DRs in SPES for all these patients separately, in a similar way as in [17]. This was done for all types of DRs and only DRs including HFOs (R, SR and SRF). Significant differences between sensitivity and specificity of both instances were tested using Wilcoxon Signed Rank test. The application of SPES is described in section 2.2.1. The consensus of the two scorers was again used as ground truth for the DRs. Table 3.1 shows the representation of different DR types inside and outside the SOZ.

3.2.3 Training and testing of CNN

Secondly, we trained a CNN for classifying SPES images on 'inside SOZ' or 'outside SOZ'. It was expected that specificity of DRs would increase when only taking DRs including HFOs into account (see Section 1.3). If this was indeed the case and sensitivity did not decrease, we would decide to only use the SPES images that showed DRs including HFOs for training and testing purposes.

Table 3.1: Representation of different DR types inside and outside SOZ

	S	R	SR	SRF	Total
Inside SOZ	60	6	116	18	200
Outside SOZ	570	121	431	44	1166

S: spike. R: ripple. SR: spike-ripple. SRF: spike-ripple-fast ripple.

The structure of the feature extraction compartment in Figure 2.3 was reused for training this CNN. Also the loss function and the adaptive learning rate method were the same and a balanced class-weight was used, as we again had a class imbalance. As the number of input images drastically decreased compared to section 2.2.4, we had to prevent the network from overfitting. Therefore, we added another 2x2 max-pooling layer to the network, followed by a 2 node, fully connected, output layer with softmax activation. Also, we trained the network using 'early stopping', meaning that the network would stop training after the loss of the test set shows no decline within 10 epochs, with a maximum of 80 epochs. The test set consisted of 10% of the total data, randomly chosen. Batch size was set to 32. We evaluated the network by calculating sensitivity and precision of the prediction of the test set to the DRs originating from inside the SOZ.

3.3 Results

3.3.1 Statistical relation to SOZ

The median number of electrodes inside SOZ was between 6 and 7 [1-15]. The DRs including HFOs showed a higher specificity for the SOZ compared to all types of DRs (mean 0.50 vs. 0.27, $p = 0.008$) and no significant difference in sensitivity (mean 0.75 vs. 0.87, $p = 0.25$). Therefore, we included only images containing R, SR or SRF for training and testing the CNN.

3.3.2 Detection of pathological DRs

Figure 3.1 shows the learning curve of the training and test set. Table 3.3 shows the confusion matrix of the test set. Sensitivity was 0.63 and precision 0.29. For the train set, the values for sensitivity and precision were 0.71 and 0.39, respectively.

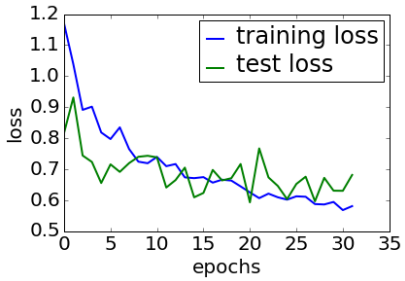


Figure 3.1: Learning curve of the CNN for detecting SOZ related DRs. Shown is the loss after each epoch of training for both train and test set (see Equation 5 in Appendix A). Training of the model stopped after the test loss had stopped declining within 10 epochs.

Table 3.2: Confusion matrix CNN for detecting SOZ on train set

		Predicted		
		Inside SOZ	Outside SOZ	Total
True	Inside SOZ	102	22	124
	Outside SOZ	162	376	538
	Total	264	398	662

The correctly predicted samples are on the diagonal, shown in gray.

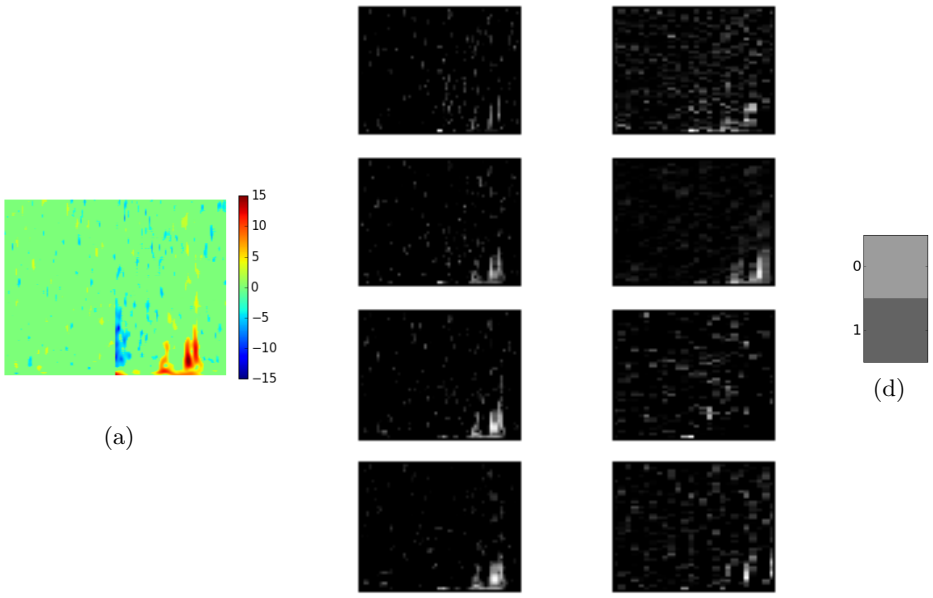
Table 3.3: Confusion matrix CNN for detecting SOZ on test set

		Predicted		
		Inside SOZ	Outside SOZ	Total
True	Inside SOZ	9	7	16
	Outside SOZ	23	35	58
	Total	32	42	74

The correctly predicted samples are on the diagonal, shown in gray.

3.3.3 Feature extraction

In an attempt to learn from the feature extraction that is learned by the model, the activations of some examples were reviewed. Figure 3.2 shows an example of the activations of the network to an input image.



(a) (b) (c) (d)

Figure 3.2: Example of activations of CNN for relating DRs to SOZ. (a): input matrix. (b): Feature maps of first convolutional layer. (c): Feature maps of second convolutional layer. (d): Softmax output. Both (b) and (c) are grayscale images scaled on its min and max values. These layers represent the features extraction of the model, meaning that they show what the model deems important in the input image for classification. In (d), 0 is outside and 1 is inside SOZ. Here, the grayscale ranges from 0 to 100%. In this example, the network is 61% sure the label is outside SOZ. The input image was part of the train set.

3.4 Discussion

We trained a CNN with the goal to see whether a DL model could succeed in extracting features from TF-SPES images that are different for physiological and pathological responses.

3.4.1 Performance of model

The balanced class-weight again resulted in an imbalance in sensitivity and precision. It can be argued that high sensitivity to DRs that originate from inside SOZ is desired over high specificity. In the context of epilepsy surgery, resecting a larger area of cortex than strictly necessary for good outcome is probably preferred over resecting not enough cortex for the patient to become seizure free. That is, as long as the resected area is not part of eloquent cortex.

However, looking at Table 3.3 and the achieved sensitivity and precision, we can conclude that the CNN does not perform well enough to fulfill that purpose. Apparently, the network

did not succeed in finding features to distinguish between inside SOZ DRs and outside SOZ DRs. This could be due to the relatively small amount of data available for training, but it is also possible that the origin of DRs simply cannot be determined from TF-SPES images, as was also the conclusion of [17]. Also when reviewing the different images, no clear differences can be distinguished by human eye. Figure 3.3 gives some examples of this.

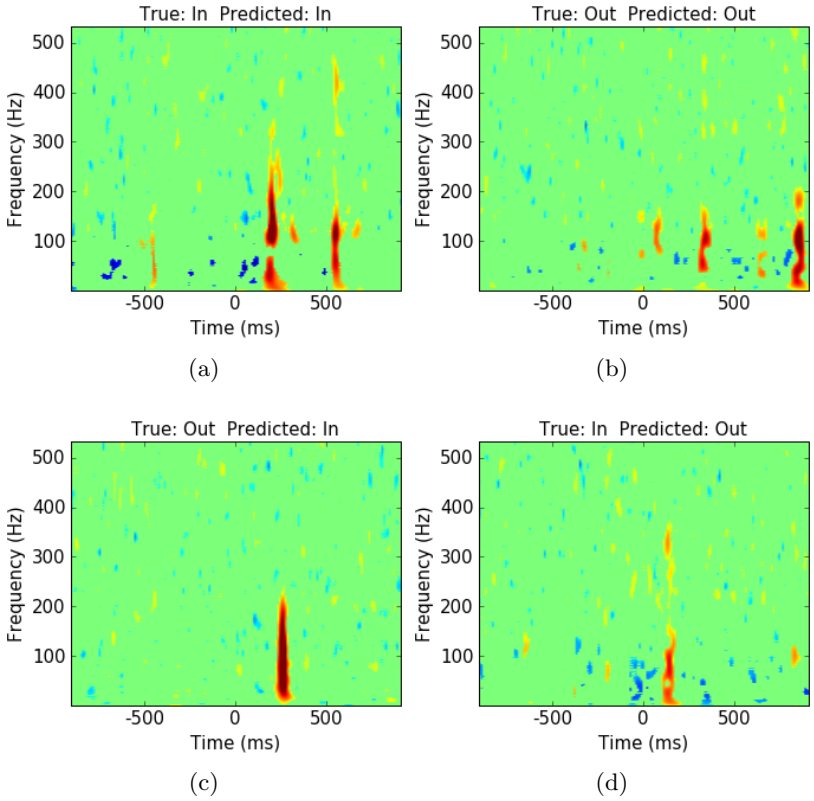


Figure 3.3: Examples of DRs originating from inside (in) and outside (out) SOZ. All images originate from the train set.

3.4.2 Feature extraction

Looking at the activations of an input image in Figure 3.2, it is remarkable that, contrary to the CNN trained for scoring DRs, none of the nodes in the first convolutional layer seems to have learned to focus on the suppression in the input image. This suggests that whether suppression occurs after stimulation is not relevant for distinguishing between pathological and physiological responses. However, this is in contradiction with the conclusions of Jacobs et al., who found that decrease in the high frequency band after a spontaneous spike is a promising marker to identify SOZ [32]. Perhaps, the performance of the model is too bad to draw a conclusion on the relevance of suppression after stimulation.

3.4.3 Methodological aspects

SOZ and pathological responses

Despite its correlation to SOZ, a large part from the DRs originate from outside SOZ. To learn more about these DRs, it could be assessed whether they were observed in eloquent cortex. Alternatively, these DRs could be an indication of the early seizure spread. The latter would mean that not all DRs outside SOZ are physiological, as was assumed in this research.

It has to be noted that in this research, we focused on relating separate DRs to SOZ. However, defining the SOZ is based on electrodes, not on DRs. It is not necessarily true that every DR observed in a SOZ electrode also has to be a SOZ related DR, which is what we assumed here. Therefore, it could be useful to take all observed DRs on one response electrode together in future analysis. It could be the case that the observation of one specific SOZ related DR is enough to appoint that response electrode to the SOZ. The distinction, however, might also be found in the frequency of the occurrence of certain DRs.

Model considerations

We decided to reuse a large part of the structure of the CNN discussed previously. We assumed that this cross-validated structure would also suffice for the purpose of relating the DRs to SOZ as no new input images were used. Thus we avoided the need for cross-validation of the network. Because we did not know whether the distinction between inside and outside SOZ could be based on the features that the network had learned to extract previously, we re-initialized the weights of the network and had them all retrained. This resulted in 1,262 trainable parameters. We used 'early stopping' to prevent the network from overfitting on the train set. Figure 3.1 shows that 32 epochs were trained and that the model had problems converging on the test set. We decided to stop training after the model had not improved on the test set after a relatively large number of 10 epochs. We chose 10 as the test loss had proven not to decrease very smoothly. It has to be noted that the number of epochs on which the model was trained can be considered as a hyper-parameter, which was optimized on the test set, thus not distinguishing between a cross-validation and test set.

Data selection

In accordance with [16] and [33], we saw an increased specificity for the SOZ when only taking HFOs and spikes including HFOs into account in the analysis. By excluding the spikes without HFOs in the analysis, we ignore the possibility that also in these images there is some activity present that can distinguish between S inside SOZ and S outside SOZ. However, we find it more likely that, if this activity is present, it will be more pronounced in the high frequency bands. That is, assumed that activity in the spike frequency band in the TF-SPES images in fact has the shape of a spike in the time domain and is not merely beta or gamma activity. This is further discussed in section 4.1.

It can be concluded from Table 3.1 that when an isolated ripple is observed, it is much more likely to originate from outside SOZ than the other types of DRs. Although it is not completely clear whether it is beneficial to distinguish between separate HFOs and HFOs that are superimposed on spikes [13], Jacobs et al. showed that HFOs more often co-occur with spikes inside SOZ [33]. Therefore, it should be considered to also leave the R scored images out of the analysis.

Although the network did not succeed in relating DRs to SOZ in the TF-SPES images, it is possible that there could be found some useful features in the raw time data, that are lost in constructing the images. Therefore, it is recommended to train a network on the raw time data of SPES. We discuss this option further in section 4.2.

Additionally, we decided to leave out ERs in our analysis. However, Mouthaan et al. show that also ERs are strongly related to SOZ and seizure propagation [34]. It should be taken into consideration to include both ERs and DRs in future analysis.

Furthermore, it is possible that the ten separate stimulations contain useful information which is lost in the averaging. A way to use this information is to create input matrices with 10 channels, each of which is one stimulation. However, it is not very obvious how the network would have to deal with the stochastic aspect of the responses. For example, it is not likely that a difference in response inside SOZ compared to outside SOZ will come from one specific stimulation. If the separate stimulations contains useful information, it is expected that it will be in the variance of the responses, which is hard for a neural network to extract out of the input data.

SOZ as ground truth

We only had one expert looking at the seizure onset, whereas the experts themselves report that there exists a large inter-rater variability in determining SOZ. Moreover, it is unclear whether SOZ is the best indication of the EZ. Lüders et al. mention that the EZ is often more extensive than the SOZ [6]. On the other hand, Jacobs et al. conclude that incomplete removal of SOZ does not necessarily result in bad outcome [14]. According to Valentín et al., the goal of SPES is to identify epileptogenic cortex, which is very similar to the EZ (see section 1.2). Currently, we use the SOZ as an indication of the EZ and in this study, we tried to let SPES be an indication of the SOZ. However, one could argue that it would make more sense to relate the responses to SPES directly to the EZ. We will elaborate on this in section 4.2.

3.4.4 Conclusion

The model did not achieve satisfactory results in distinguishing between DRs originating from inside and outside SOZ. This could be due to the small amount of data. It could also be possible that we wrongly assumed to observe all pathological DRs inside the SOZ. Alternatively, it is possible that the distinction simply cannot be based on the TF-SPES images. More research is needed.

General discussion

4.1 Methodological aspects

A general impediment for the analysis of SPES is the presence of spontaneous interictal activity. The distinction between IDs and DRs is important, as DRs are better related to SOZ [9, 35], but if IDs are present, it may be hard to tell the difference between spontaneous activity and an evoked response. Due to the averaging of 10 power spectra in the construction of ERSPs, the evoked responses with approximately the same latency often show a larger power increase compared to the spontaneous activity [16]. Nonetheless, it remains something that has to be dealt with.

Machine learning usually needs a large amount of data to learn from, before it can reliably be applied to new data. Therefore, we chose to take the TF-SPES images of all patients together, thus ignoring the inter-patient variability. Also, we ignored the difference in brain regions from which the responses originate. Valentín describes different kinds of responses for frontal and temporal lobe epilepsies [11]. Furthermore, it is possible that different etiologies of epilepsy, e.g. a tumor or focal cortical dysplasia, also may have different responses to SPES. As the number of patients was limited in this research, it is possible that these inter-patient differences have influenced the performance of the models. A way to overcome this problem is to take the brain area, etiology, and possibly other patient specific data, such as age, into account in future analyses. However, each subgroup again needs a large number of examples for the model to learn the differences in responses.

Urrestarazu et al. concluded that HFOs sometimes occur as filter artifacts of spikes [36]. Although this only concerned a small minority of the observed HFOs, it appears that the time-frequency analysis of HFOs is not the optimal way to distinguish between real oscillations and artifacts [37]. On the other hand, Van 't Klooster et al. argue that regardless of the origin of an observed HFO, it is statistically related to the seizure onset zone and therefore clinically relevant [16]. Also, the time-frequency analysis has shown to be faster than visual analysis in time domain and leads to equal or better inter-observer agreement [17]. However, Jacobs et al. conclude that replacing HFO analysis in time domain by time-frequency analysis in a clinical setting is not recommended, as the latter does not seem to be able to completely represent a spike with superimposed HFOs [32]. Indeed, it is sometimes observed that what appears to be a spike in the TF-SPES image, in the time domain actually is beta or gamma

activity rather than an actual spike. The difference between both phenomenons may have clinical value, but is ignored in the time-frequency analysis of SPES.

4.2 Future directions

Some possible improvement for the models trained in this research have been proposed in the previous sections. In this section, we want to propose a different approach to the problem.

As was briefly mentioned in section 1.1, Bartolomei et al. argue that the model of an epileptic focus is inaccurate, as often seizure onset is accompanied by simultaneous involvement of several distributed brain regions [5]. Epilepsy seems to be a network disease, rather than a structural disease. Consequently, removal of one specific focus does not necessarily result in good outcome. Although it remains uncertain how to deal with this in the context of epilepsy surgery, it may be argued that SPES functions as an even more reliable indicator for the cortical imbalance than the SOZ. We argue that there are two, equally important, approaches to further elaborate on this.

The first is to construct computational models in order to better understand the dynamics of an epileptogenic network and its responses to SPES, as is currently done by Hebbink et al. [38]. If these models are able to represent human cortex well enough, it offers a very good framework to define which part of the epileptogenic network necessarily has to be disconnected to produce seizure freedom and how this could be found by stimulations to the network.

The second, phenomenological, approach is to train a CNN to directly relate SPES responses to patient outcome. This can be done by selecting a large amount of patients with good outcome after surgery (Engel class I) and distinguish between resected and not resected area. Based on the aforementioned, it is recommended to use the time data instead of the time-frequency images as input. The ground truth labels of the responses are set to 'needs to be resected for good outcome' and 'does not need to be resected for good outcome'. Similar to the 'inside SOZ' and 'outside SOZ' distinction, this is a label made on electrode level, not on response level. Therefore, we recommend to combine the responses of a response electrode to all stimulation locations into one input. The time data are the columns of the input matrix and the rows represent different stimulations. The problem that every patient has a different number of stimulations could be solved by only looking at the 10 stimulations that showed the largest number of ERs. This is currently a subject of research of Hebbink et al. The responses to the 10 stimulations that are given to each stimulation location could be averaged in time domain or could be used as separate input 'channels', thus creating a 3 dimensional input matrix.

An example of a CNN trained on raw EEG data is given in [39] and [25]. A difference with these examples is that the classifications are on the entire EEG, i.e. outcome or sex of patient, rather than on specific electrodes, i.e. EZ electrode or not. However, if the input matrices are constructed as suggested here, similar techniques could be applied.

One remaining problem with this approach is that the resected area in patients with good outcome not necessarily consists of the 'least amount of cortex' that has to be resected to achieve good outcome.

4.3 Relation to other work

Dümpelmann et al. introduced a radial basis function neural network for automatic detection of ripples [40]. Although it is called a neural network, it is actually more similar to the SVM as used in this research. The used features were estimates of the energy and signal length of the time data. The third feature was based on the frequency domain. Their achieved sensitivity, specificity and kappas were all below 0.5.

Jrad et al. were the first to build a detector for the whole high frequency band, ranging from gamma to fast ripple [41]. They made use of Gabor atoms to decompose the time signal into its frequency components. This is very similar to the wavelet transform used in this research. Their detection of ROIs in this decomposition was based on the root mean square energy of all HFO frequency bands. From the detected ROIs, they extracted features based on a certain energy ratio for all different bands and the duration of the ROI. The features were used as an input for an SVM with radial basis function, similar to the one used in this research. In their artifact class, they also included high frequency activity that was due to filter artifacts of a spike. Compared to our results, they achieved high specificity, which could be due to the fact that they did not use the 'empty' signals, i.e. signals without ROIs, as an input for the SVM. Also, over 1000 examples for each HFO class was available for training and testing purposes. In contrast to this research, Jrad et al. did not include mixed classes (e.g. SRF) in their analysis. Thus it is unclear how HFOs co-occurring with spikes were treated.

To the best of our knowledge, we were the first to apply deep learning for the detection of HFOs, and the first to apply machine learning in the context of SPES.

4.4 Clinical context

The definition of EZ as "the minimum amount of cortex that must be resected to produce seizure freedom" is only a theoretical one. This means that in clinical practice, we can only in hindsight know whether the EZ has in fact been resected. Therefore, all available technologies are currently used in the presurgical evaluation to gather as much information as possible.

In the UMC Utrecht, the decision for epilepsy surgery commonly is a weighted sum of the intracranially observed SOZ, IDs (irritative zone), spontaneous HFOs, and finally, DRs to SPES. The latter is often used merely as a confirmation for the existing hypothesis. Other things that are taken into account are the semiology of seizures (symptomatogenic zone) and the functionality found with ESM. If a seizure is induced during any of these stimulations, or if the patient reports to have the sensation of a seizure, this is also considered to be useful information. Furthermore, if a lesion is observed on the MRI (epileptogenic lesion) and/or an ictal SPECT is done, this is all part of the equation. In fact, an observed dysplasia might in clinical practice be the biggest indicator for the EZ. If a clear lesion is observed, it is only assessed whether the iEEG is in concordance with it. However, as was discussed previously, the removal of a specific focus does not necessarily result in good outcome. This shows that despite all the different modern technologies, it is very difficult to bring our theoretical

knowledge of epilepsy into the clinical practice of surgery.

The fundamental idea behind SPES is that it can indicate the EZ more efficiently than by waiting for spontaneous seizures, IDs or HFOs. If we can gain more knowledge on how the responses to SPES should be interpreted, this would mean that waiting for spontaneous activity is no longer of added value. Then, if both the protocol and the analysis of SPES are further optimized, for example by the research of Hebbink et al. and the methods suggested in this research, SPES might be applied intra-operatively. Provided that ESM is feasible during surgery or not needed at all, this would entirely eliminate the need for a prolonged ECoG recording on the IEMU.

This research is again a step towards this efficient and automated delineation of focal epilepsy with the use of SPES.

4.5 Additional notes

Machine learning methods have been applied for over 60 years in training computers to outperform humans at a certain task. Initially, these tasks mainly were focused on playing games. For example, already in the 1950s, IBM worked on a computer that taught itself how to play the game of checkers [42]. It is well known that many years later, in 1997, IBM's computer Deep Blue defeated the world champion Garry Kasparov in an official match of chess. The reason why games were such a popular subject for machine learning is explained by Samuel in the following way: "A game provides a convenient vehicle for such study as contrasted with a problem taken from life, since many of the complications of detail are removed"[42]. This certainly makes sense, but with the invention of new ML techniques and the increasing amount of data available, ML gradually also found its way into the medical world. As computational power strongly increased over the years, especially DL grew in popularity. In medical imaging, a lot of research has already been done with the use of DL [22, 43]. In fact, one of the world's biggest experts on DL, Geoffrey Hinton, recently said that people should stop training radiologists as DL will very soon outperform humans in this task. I foresee endless possibilities for neural networks in other medical fields as well, not only in making the analysis of clinical data more efficient, as we did in the first part of this research, but also in finding new features in that data upon which new types of analyses can be based, as was attempted in the second part of this research. However, the transition from ML in research to its application in clinical practice seems to be problematic. Obviously, basing clinical decisions blindly on the output of a computer must be done with care. Thus, this transition needs to be guided. I strongly believe that the clinical physician is ought to play a crucial role in the implementation of these cutting edge ML techniques into clinical practice. On the one hand, the clinical physician knows the possibilities of the technology and, even more important, its limitations. On the other hand, the clinical physician fully understands the different aspects of the clinical problem and knows what is required to solve it.

Bibliography

[1] Anthony K Ngugi, Christian Bottomley, Immo Kleinschmidt, Josemir W Sander, and Charles R Newton. Estimation of the burden of active and life-time epilepsy : A meta-analytic approach. *Epilepsia*, 51(5):883–890, 2010. doi: 10.1111/j.1528-1167.2009.02481.x.

[2] J.W.A.S. Sander. Some aspects of prognosis in the epilepsies: a review. *Epilepsia*, 34 (6):1007, 1993. ISSN 0013-9580.

[3] Kristina Malmgren and Anna Edelvik. Long-term outcomes of surgical treatment for epilepsy in adults with regard to seizures, antiepileptic drug treatment and employment. *Seizure*, 44:217–224, 2017. ISSN 15322688. doi: 10.1016/j.seizure.2016.10.015. URL `http://dx.doi.org/10.1016/j.seizure.2016.10.015`.

[4] Nathalie Jetté, Josemir W. Sander, and Mark R. Keezer. Surgical treatment for epilepsy: the potential gap between evidence and practice. *The Lancet Neurology*, 15(9):982–994, 2016. ISSN 14744465. doi: 10.1016/S1474-4422(16)30127-2. URL `http://dx.doi.org/10.1016/S1474-4422(16)30127-2`.

[5] Fabrice Bartolomei, Stanislas Lagarde, Fabrice Wendling, Aileen McGonigal, Viktor Jirsa, Maxime Guye, and Christian Bénar. Defining epileptogenic networks: Contribution of SEEG and signal analysis. *Epilepsia*, 58(7):1131–1147, 2017. ISSN 15281167. doi: 10.1111/epi.13791.

[6] Hans O Lüders, Imad Najm, Dileep Nair, and Peter Widdess-walsh. The epileptogenic zone : general principles. *Epileptic Disorders*, 8(August):1–9, 2006.

[7] Prasanna Jayakar, Jean Gotman, A. Simon Harvey, André Palmini, Laura Tassi, Donald Schomer, Francois Dubeau, Fabrice Bartolomei, Alice Yu, Pavel Kršek, Demetrios Velis, and Philippe Kahane. Diagnostic utility of invasive EEG for epilepsy surgery: Indications, modalities, and techniques. *Epilepsia*, 57(11):1735–1747, 2016. ISSN 15281167. doi: 10.1111/epi.13515.

[8] A Valentín, M Anderson, G Alarcón, J J García Seoane, R Selway, C D Binnie, and C E Polkey. Responses to single pulse electrical stimulation identify epileptogenesis in the human brain in vivo. *Brain : a journal of neurology*, 125(Pt 8):1709–1718, 2002. ISSN 0006-8950. doi: 10.1093/brain/awf187.

[9] Antonio Valentín, Gonzalo Alarcón, Mrinalini Honavar, Jorge J. García Seoane, Richard P. Selway, Charles E. Polkey, and Colin D. Binnie. Single pulse electrical stimulation for identification of structural abnormalities and prediction of seizure outcome after epilepsy surgery: A prospective study. *Lancet Neurology*, 4(11):718–726, 2005. ISSN 14744422. doi: 10.1016/S1474-4422(05)70200-3.

[10] A. Valentín, G. Alarcón, J. J. García-Seoane, M. E. Lacruz, S. D. Nayak, M. Honavar, R. P. Selway, C. D. Binnie, and C. E. Polkey. Single-pulse electrical stimulation identifies epileptogenic frontal cortex in the human brain. *Neurology*, 65(3):426–435, 2005. ISSN 00283878. doi: 10.1212/01.wnl.0000171340.73078.c1.

[11] Antonio Valentín. Single-pulse electrical stimulation. In *Introduction to Epilepsy*, pages 452–455. 2012. ISBN 9781139103992.

[12] Antonio Valentín, Gonzalo Alarcón, Sally F. Barrington, Jorge J. García Seoane, María C. Martín-Miguel, Richard P. Selway, and Michalis Koutroumanidis. Interictal estimation of intracranial seizure onset in temporal lobe epilepsy. *Clinical Neurophysiology*, 125(2):231–238, 2014. ISSN 13882457. doi: 10.1016/j.clinph.2013.07.008.

[13] Maeike Zijlmans, Premysl Jiruska, Rina Zelmann, Frans S S Leijten, John G R Jefferys, and Jean Gotman. High-Frequency Oscillations as a New Biomarker in Epilepsy. *Annals of Neurology*, 71:169–178, 2012. doi: 10.1002/ana.22548.

[14] Julia Jacobs, Maeike Zijlmans, Rina Zelmann, and Jeffrey Hall. High-Frequency Electroencephalographic Oscillations Correlate With Outcome of Epilepsy Surgery. *Annals of Neurology*, 67:209–220, 2010. doi: 10.1002/ana.21847.

[15] John D. Rolston, Nealen G. Laxpati, Claire Anne Gutekunst, Steve M. Potter, and Robert E. Gross. Spontaneous and evoked high-frequency oscillations in the tetanus toxin model of epilepsy. *Epilepsia*, 51(11):2289–2296, 2010. ISSN 00139580. doi: 10.1111/j.1528-1167.2010.02753.x.

[16] Maryse A. Van 't Klooster, Maeike Zijlmans, Frans S S Leijten, Cyrille H. Ferrier, Michel J A M Van Putten, and Geertjan J M Huiskamp. Time-frequency analysis of single pulse electrical stimulation to assist delineation of epileptogenic cortex. *Brain*, 134(10):2855–2866, 2011. ISSN 14602156. doi: 10.1093/brain/awr211.

[17] M. A. van 't Klooster, N. E C van Klink, D. van Blooijs, C. H. Ferrier, K. P J Braun, F. S S Leijten, G. J M Huiskamp, and M. Zijlmans. Evoked versus spontaneous high frequency oscillations in the chronic electrocorticogram in focal epilepsy. *Clinical Neurophysiology*, 128(5):858–866, 2017. ISSN 18728952. doi: 10.1016/j.clinph.2017.01.017. URL `http://dx.doi.org/10.1016/j.clinph.2017.01.017`.

[18] Tom M. Mitchell. *Machine learning in ecosystem informatics and sustainability*. McGraw-Hill Science/Engineering/Math, 1997. ISBN 9781577354260. doi: 10.1007/978-3-540-75488-6_2.

[19] V Vapnik. *The Nature of statistical Learning Theory*. 1995. ISBN 9781475724424. doi: 10.1007/978-1-4757-2440-0. URL `http://infoscience.epfl.ch/record/82790/files/com02-04.pdf`.

[20] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015. ISSN 14764687. doi: 10.1038/nature14539.

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998. ISSN 00189219. doi: 10.1109/5.726791.

[22] Kenji Suzuki. Pixel-based machine learning in medical imaging. *International Journal of Biomedical Imaging*, 2012, 2012. ISSN 16874188. doi: 10.1155/2012/792079.

[23] Shih Chung B Lo, Heang Ping Chan, Jyh Shyan Lin, Huai Li, Matthew T. Freedman, and Seong K. Mun. Artificial convolution neural network for medical image pattern recognition. *Neural Networks*, 8(7-8):1201–1214, 1995. ISSN 08936080. doi: 10.1016/0893-6080(95)00061-5.

[24] Alex Krizhevsky, Ilya Sutskever, and Hinton Geoffrey E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS2012)*, pages 1–9, 2012. ISSN 10495258. doi: 10.1109/5.726791. URL `https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`.

[25] Michel J. A. M. Van Putten, Sebastian Olbrich, and Martijn Arns. Predicting sex from brain rhythms with deep learning. *Scientific Reports*, 8(1):3069, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-21495-7. URL `http://www.nature.com/articles/s41598-018-21495-7`.

[26] Arnaud Delorme and Scott Makeig. EEGLAB: an open sorce toolbox for analysis of single-trail EEG dynamics including independent component anlaysis. *Journal of Neuroscience Methods*, 134:9–21, 2004. ISSN 0165-0270. doi: 10.1016/j.jneumeth.2003.10.009.

[27] Brian S Everitt. *The analysis of contingency tables*. CRC Press, 1992.

[28] E L Allwein, R Schapire, and Y Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *The Journal of Machine Learning*, 1:113–141, 2001. ISSN 15324435. doi: 10.1162/15324430152733133.

[29] François Chollet et al. Keras. `https://github.com/keras-team/keras`, 2015.

[30] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[31] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. ISSN 10769757. doi: 10.1613/jair.953.

[32] Julia Jacobs, Christina Vogt, Pierre LeVan, Rina Zelmann, Jean Gotman, and Katsuhiro Kobayashi. The identification of distinct high-frequency oscillations during spikes delineates the seizure onset zone better than high-frequency spectral power changes. *Clinical Neurophysiology*, 127(1):129–142, 2016. ISSN 18728952. doi: 10.1016/j.clinph.2015.04.053. URL `http://dx.doi.org/10.1016/j.clinph.2015.04.053`.

[33] Julia Jacobs, Pierre LeVan, Rahul Chander, Jeffery Hall, François Dubeau, and Jean Gotman. Interictal high-frequency oscillations (80-500 Hz) are an indicator of seizure onset areas independent of spikes in the human epileptic brain. *Epilepsia*, 49(11):1893–1907, 2008. ISSN 00139580. doi: 10.1111/j.1528-1167.2008.01656.x.

[34] B. E. Mouthaan, M. A. Van't Klooster, D. Keizer, G. J. Hebbink, F. S S Leijten, C. H. Ferrier, M. J A M Van Putten, M. Zijlmans, and G. J M Huiskamp. Single Pulse Electrical Stimulation to identify epileptogenic cortex: Clinical information obtained from early evoked responses. *Clinical Neurophysiology*, 127(2):1088–1098, 2016. ISSN 18728952. doi: 10.1016/j.clinph.2015.07.031.

[35] Dinesh Nayak, Antonio Valentín, Richard P. Selway, and Gonzalo Alarcón. Can single pulse electrical stimulation provoke responses similar to spontaneous interictal epileptiform discharges? *Clinical Neurophysiology*, 125(7):1306–1311, 2014. ISSN 18728952. doi: 10.1016/j.clinph.2013.11.019. URL `http://dx.doi.org/10.1016/j.clinph.2013.11.019`.

[36] Elena Urrestarazu, Rahul Chander, François Dubeau, and Jean Gotman. Interictal high-frequency oscillations (10-500 Hz) in the intracerebral EEG of epileptic patients. *Brain*, 130(9):2354–2366, 2007. ISSN 00068950. doi: 10.1093/brain/awm149.

[37] Mina Amiri, Jean Marc Lina, Francesca Pizzo, and Jean Gotman. High Frequency Oscillations and spikes: Separating real HFOs from false oscillations. *Clinical Neurophysiology*, 127(1):187–196, 2016. ISSN 18728952. doi: 10.1016/j.clinph.2015.04.290. URL `http://dx.doi.org/10.1016/j.clinph.2015.04.290`.

[38] Jurgen Hebbink, Hil Meijer, Geertjan Huiskamp, Stephan van Gils, and Frans Leijten. Phenomenological network models: Lessons for epilepsy surgery. *Epilepsia*, 58(10):e147–e151, 2017. ISSN 15281167. doi: 10.1111/epi.13861.

[39] Michel J. A. M. van Putten, Jeannette Hofmeijer, Barry J. Ruijter, and Marleen C. Tjepkema-Cloostermans. Deep learning for outcome prediction of postanoxic coma. In Hannu Eskola, Outi Väisänen, Jari Viik, and Jari Hyttinen, editors, *EMBEC & NBC 2017*, pages 506–509, Singapore, 2018. Springer Singapore. ISBN 978-981-10-5122-7.

[40] Matthias Dümpelmann, Julia Jacobs, Karolin Kerber, and Andreas Schulze-Bonhage. Automatic 80-250Hz "ripple" high frequency oscillation detection in invasive subdural grid and strip recordings in epilepsy by a radial basis function neural network. *Clinical Neurophysiology*, 123(9):1721–1731, 2012. ISSN 13882457. doi: 10.1016/j.clinph.2012.02.072.

[41] Nisrine Jrad, Amar Kachenoura, Isabelle Merlet, Fabrice Bartolomei, Anca Nica, Arnaud Biraben, and Fabrice Wendling. Automatic detection and classification of High Frequency Oscillations in depth-EEG signals. *IEEE Transactions on Biomedical Engineering*, 64(9):1–1, 2016. ISSN 0018-9294. doi: 10.1109/TBME.2016.2633391. URL `http://ieeexplore.ieee.org/document/7762043/`.

[42] A. L. Samuel. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. ISSN 0018-8646. doi: 10.1147/rd.33.0210. URL `http://ieeexplore.ieee.org/document/5392560/`.

[43] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42(December 2012):60–88, 2017. ISSN 13618423. doi: 10.1016/j.media.2017.07.005.

[44] JP Vert, K Tsuda, and B Schölkopf. A primer on kernel methods. *Kernel Methods in Computational Biology*, (1992):35–70, 2004. ISSN 1098-6596. doi: 10.1017/CBO9781107415324.004.

[45] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. ISSN 15337928. doi: 10.1214/12-AOS1000.

[46] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. ISSN 00280836. doi: 10.1038/323533a0.

Appendices

A Mathematics of ML models

A.1 Class-weights

The balanced class-weight was introduced to deal with the class imbalance. Balanced class-weight means that, during training, the loss is weighted in the following way:

$$w_k = \frac{n}{n_k * K}, \tag{1}$$

where w_k is the weight for class k, n is the total number of samples, n_k is the number of samples in class k and K is the total number of classes.

A.2 SVM

Objective function

For the SVM, each binary classifier is trained by minimizing the objective function:

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C * L, \tag{2}$$

where $\frac{1}{2}\|\boldsymbol{\theta}\|^2$ is a regularization term, which should prevent the classifier from overfitting, and L is the Hinge loss:

$$L = \sum_{i=1}^{n} w_{i,k} * \max[0,\, 1 - t_{i,k} * y_i(\mathbf{x_i}, \boldsymbol{\theta})], \tag{3}$$

with $\boldsymbol{\theta}$ being the vector of trainable parameters and $\mathbf{x_i}$ the feature vector. $w_{i,k}$ is class-weight (see Equation 1) and $t_{i,k}$ is target value, which equals [1,-1] for the classes of the binary classifier and 0 otherwise. n is the total number of samples.

Gaussian radial basis function kernel

The Gaussian radial basis function kernel is included in y_i and makes it a non-linear function [44].

A.3 CNN

Dropout layers

Dropout layers randomly exclude each input with a certain chance p in learning phase [45]. In testing phase, the parameters of the next layer are weighted with factor p.

Objective function

For the CNN, the objective function was as follows:

$$\min_{\mathbf{\Theta}} L \tag{4}$$

where cross-entropy was chosen as the loss function:

$$L = -\sum_{i=1}^{n} \mathbf{t_i} \circ \mathbf{w} \cdot \log \mathbf{y_i} \tag{5}$$

Here, $\mathbf{w}$ is the class-weight vector and $\mathbf{t_i}$ is the target vector of the i^{th} sample: $\mathbf{t_i} = [t_1, t_2, ..., t_K]$, where K the number of classes. The value of t_k is 1 for the class of sample i and 0 otherwise. This is called a one-hot representation. n is the total number of samples in the learning batch, which was set to 250, and $\circ$ depicts element-wise multiplication. $\mathbf{y_i}$ is the output vector of sample i:

$$\mathbf{y_i} = g(\mathbf{\Theta a_i}) \tag{6}$$

where $\mathbf{\Theta}$ is the parameter matrix of the output layer and g its non-linear activation function. $\mathbf{a}$ is the output vector of the previous layer, which is calculated in a similar way as $\mathbf{y}$.

Backpropagation

The algorithm minimizes loss by updating the parameters $\mathbf{\Theta}$ of layer l with gradient descent after every iteration:

$$\mathbf{\Theta}^{(l)} = \mathbf{\Theta}^{(l)} - \eta * \frac{\partial L}{\partial \mathbf{\Theta}^{(l)}} \tag{7}$$

where η is the learning rate and $\frac{\partial L}{\partial \mathbf{\Theta}^{(l)}}$ is calculated using the backpropagation algorithm [46].

Activation functions

The activation function for ReLU is as follows:

$$g(\mathbf{x}) = \max(\mathbf{x}, 0). \tag{8}$$

The value of each output node, with Softmax activation, was calculated in the following way:

$$g(\mathbf{x})_k = \frac{e^{\mathbf{x}_k}}{\sum_{j=1}^{K} e^{\mathbf{x}_j}}, \tag{9}$$

where K is again the number of classes and x_k is the k^{th} value of the input vector x.

B Verantwoording

Het loopt alweer richting het einde, mijn afstudeerstage. Ik heb het afgelopen jaar veel geleerd en ben in mijn persoonlijke ontwikkeling weer verder gegroeid. Ik denk dat ik met recht kan zeggen dat ik me nu echt een Technisch Geneeskundige voel.

In dit verslag zal ik ingaan op de persoonlijke leerdoelen die ik in de beginfase van mijn stage heb opgesteld en hoe ik me met deze leerdoelen heb beziggehouden. Ook zal ik de activiteiten noemen welke niet direct gerelateerd zijn aan mijn leerdoelen of hebben bijgedragen aan mijn eindverslag, maar welke desalniettemin hebben bijgedragen aan mijn persoonlijke ontwikkeling.

B.1 Leerdoelen

Mijn eerste leerdoel was 'de regie nemen in mijn opdracht'. Hiermee bedoelde ik dat ik niet te zelfstandig wilde werken en actief mensen bij mijn opdracht betrekken die mij verder zouden kunnen helpen. Of, zoals het in het stageboek staat geformuleerd: Van de student Technische Geneeskunde wordt namelijk verwacht dat hij goed onderscheid leert te maken tussen wat hij (beter) zelf kan op- en uitzoeken en wat juist (beter) vanuit een samenwerking met andere collegas/professionals tot stand komt. Ik denk dat ik daar tevreden over kan zijn. In de beginfase heb ik overleggen gehad met mensen van het Julius centrum en de BCI en uiteindelijk heb ik Christoph Brune van de UT betrokken bij vooral het machine learning gedeelte van mijn opdracht. Hij is dan ook medeauteur geworden van het abstract dat ik heb ingediend voor het congres in Washington DC in mei. Verder ben ik voornamelijk veel bij Geertjan (technologisch begeleider op afdeling) binnengelopen als ik ergens tegenaan liep of om iets inhoudelijks te bespreken. Ten slotte heb ik gedurende de stage een aantal keer een presentatie gemaakt om mijn efficint mijn voortgang te tonen en te bespreken met verschillende begeleiders. Dit dwong me daarbij ook om overzicht te creren in wat ik aan het doen was.

Het tweede doel was 'mezelf laten zien in de kliniek'. Hierbij had ik vooral het doel een zekere verantwoordelijkheid naar me toe te kunnen trekken, zodat de klinische ervaring het meekijken zou overstijgen. Dit is geslaagd op twee manieren. De eerste is door twee achtereenvolgende weken als coassistent op de afdeling (algemene Neurologie en Cerebrovasculaire Ziekten) mee te lopen. Hierdoor heb ik zelfstandig een anamnese en neurologisch lichamelijk onderzoek kunnen afnemen en heb ik de arts-assistent kunnen bijstaan met bijvoorbeeld te overleggen met medebehandelaars. De tweede manier hoe dit doel is bereikt is door een aantal avond- en nachtdiensten te draaien als laborant op de Intensieve Epilepsie Monitoring Unit (IEMU). Hierbij zijn we, samen met een verpleegkundige, verantwoordelijk voor (het in de gaten houden van) de patint, voornamelijk wanneer deze een aanval krijgt. We helpen de behandelend neurologen met het zetten van markers in het EEG. Mooi meegenomen is dat we hier ook nog een zakcentje mee verdienen.

Het laatste doel was 'bewust met mijn procesontwikkeling bezig zijn'. Dit is een leerdoel waar de intervisies en 360 graden feedback momenten natuurlijk je wel toe dwingen. Ik wilde hierbuiten echter ook mijn ontwikkeling niet uit het oog verliezen. Hiervoor heb ik

elke maand voor mezelf een maandverslag getypt over hoe het die maand was gegaan. In elk maandverslag probeerde ik kort te kijken hoe ik die maand met mijn leerdoelen was bezig geweest. Ook heb ik elke maand een meeting met mijn begeleiders op de afdeling proberen te plannen om het over mijn ontwikkeling te hebben. Deze bijeenkomsten waren in de praktijk echter meer opdracht inhoudelijk. Ook denk ik dat dit leerdoel in de laatste twee maanden een beetje is ingezakt, zoals ik ook in mijn laatste 360 graden feedback heb beschreven. Dit is absoluut iets om aan te blijven werken als ik straks aan een baan ga beginnen, zeker aangezien er dan hoogstwaarschijnlijk geen verplichte feedback momenten zullen zijn. Waarschijnlijk is het een goed idee om de persoonlijke maandverslagen voort te zetten in mijn verder carrire.

B.2 Overige activiteiten

Ten slotte zijn er gedurende mijn afstudeerstage een hoop overige activiteiten geweest die hebben bijgedragen aan mijn vaardigheid, kennis en ontwikkeling. Ik heb deze onderverdeeld in kliniek en wetenschap. 10 maanden is lang, dus het zou kunnen dat ik een paar dingen ben vergeten te noemen.

Kliniek

Ten eerste heb ik veel kunnen meekijken bij de KNF gerelateerde kliniek die in het UMC (en WKZ) gebeurt. Voorbeelden hiervan zijn de epilepsie poli, de First Seizure Clinic en klinische EEG-registraties, bijvoorbeeld na slaapdeprivatie. Bij dit laatste voorbeeld heb ik ook onder supervisie zelf EEGs kunnen beoordelen en verslaan. Ook tijdens het eerdergenoemde meelopen als coassistent op de verpleegafdeling heb ik ervaring op kunnen doen met statusvoering. Voor het leren lezen van EEGs heb ik vooral veel te danken aan de wekelijkse onderwijsmomenten van Frans (medisch begeleider). Hiervoor heb ik samen met de overige TG-studenten elke week EEGs voorbereid en af en toe ook zelf kunnen presenteren. Ik denk dat ik kan concluderen dat ik hierdoor, wat betreft het beoordelen van een EEG, behoorlijk in de richting van het niveau van een arts-assistent ben gegaan in het afgelopen jaar. Dit is overigens ook iets wat ik specifiek genoemd heb in mijn persoonlijke leerdoel 'mezelf laten zien in de kliniek'. Hier kan ik dus tevreden mee zijn. Wel moet erbij vermeld worden dat ik in de breedte van de neurologische kennis waarschijnlijk wel wat achterblijf ten opzichte van bijvoorbeeld een semiarts. Desalniettemin heb ik ook over de meer algemene neurologie veel kennis opgedaan, onder andere door de wekelijkse patintendemonstraties van de Neurologie. Ook heb ik een ochtend mee kunnen kijken op de angiokamer, waar veel neurologie gerelateerde ingrepen gebeuren, zoals het coilen van een aneurysma. Ook heb ik op de afdeling kunnen assisteren bij het doen van lumbale puncties. Voor het verbeteren van mijn anatomische kennis van het brein heb ik onder andere een hemisferotomie kunnen bijwonen (een operatie waarbij n hersenhelft volledig disfunctioneel wordt gemaakt). Ook hebben ik en mijn mede TG-studenten, onder supervisie, bij elkaar EEGs geplakt, lichamelijk onderzoek gedaan en venapuncties afgenomen. Ten slotte is er dan de kliniek die direct gerelateerd is aan de epilepsie chirurgie. Hierbij moet gedacht worden aan het bijwonen van patintbesprekingen, zoals die van de landelijke werkgroep epilepsie chirurgie, en het verdiepen in de casussen van de epilepsie chirurgie kandidaten. Ook heb ik regelmatig de gridimplantaties en

resecties op de OK kunnen bijwonen. Gedurende de periodes dat er patinten op de IEMU lagen heb ik kunnen assisteren in het mappen van de functionele gebieden en heb ik regelmatig zelfstandig Single Pulse Electrical Stimulation (SPES) uitgevoerd. Daarbij heb ik ook vaak uitgebreid naar de gemeten activiteit en aanvallen gekeken en dit besproken met Frans of Cyrille. Rondom de gridperiode heb ik ook Epitrack (een bepaalde cognitieve test) en specifieke taaltesten bij de patinten afgenomen.

Wetenschap

Wat betreft de wetenschap is denk ik het belangrijkste dat ik patinten heb gencludeerd voor de SPES-Neural Mass Model (NMM) studie waar Jurgen Hebbink als PhD'er bij betrokken is. Ik heb hiervoor de patinten gemaild en gebeld, inclusiegesprekken gevoerd en de status bijgehouden in HiX. Deze inclusie heeft niet te maken met mijn eigen onderzoek, aangezien ik alleen klinische SPES data heb gebruikt. Ook heb ik de maandelijkse SPES-NMM bijeenkomsten, met de begeleiders van Jurgen van het UMC en de UT, bijgewoond en hier mijn eigen voortgang kunnen bespreken. Dit laatste heb ik ook kunnen doen op de maandelijkse wetenschapsbesprekingen van de afdeling, waar ook regelmatig een andere student of PhD'er zijn onderzoek presenteerde en ik input kon leveren voor het vervolg ervan. Daarbij vond er een maandelijkse Journal Club onder PhD studenten plaats waar iemand elke bijeenkomst een wetenschappelijk artikel kritisch besprak met de overige aanwezigen. Hier ben ik vaak bij aanwezig geweest en in maart zal ik hier ook zelf een bijdrage aan leveren door een artikel te presenteren. Verder waren er maandelijkse RIBS praatjes, welke ik vaak heb bijgewoond. Deze praatjes gingen vaak over wat meer fundamenteel hersenwetenschappelijk onderzoek. Ook hier zal ik een maart zelf een presentatie geven over mijn onderzoek. Ten slotte heb ik tijdens mijn afstuderen een abstract ingediend voor het International Congress of Clinical Neurophysiology in Washington, waar ik in mei een poster mag presenteren, en heb ik over mijn laatste M2 stage een poster mogen presenteren op het congres van de Neurocritical Care Society in Hawaii. Deze ervaring heeft er uiteraard aan bijgedragen dat ik ernaar uit zie om na mijn afstuderen mijn carrire voort te zetten in de wetenschap. Hiervoor heb ik overigens ook meegeschreven aan een beursaanvraag voor een PhD bij de IC in het UMC Utrecht. Wie weet waar het me brengen zal?