UNIVERSITY OF TWENTE.

Enhancing interictal epileptiform discharge detection via a deep learning approach

Author

E. MEULENBRUGGE

The great enemy of truth is very often not the lie--deliberate, contrived and dishonest--but the myth--persistent, persuasive and unrealistic. Too often we hold fast to the cliches of our forebears. We subject all facts to a prefabricated set of interpretations. We enjoy the comfort of opinion without the discomfort of thought.

- John F. Kennedy

Summary

The electroencephalogram (EEG) is one of the most widely used diagnostic tools within neurology and provides valuable information about the condition of the underlying cortex in a non-invasive manner. It is predicted, that in the near future, the EEG will play an even more prominent role as it does today. This prediction is based on the sharply increasing prevalence of neurological disease as people age.
The increase of neurological diseases leads to an increased usage of the EEG and an increased burden on the already scarce personnel who need to visually analyze the EEG. Automating (a part of) this task would not only decrease this burden of the personnel but could also increase the consistency of the diagnosis by eliminating interrater variability.

In this thesis we make two contributions towards the automated analysis of the EEG by enhancing an interictal epileptiform discharge (IED) detection algorithm called SpikeNet and by validating and enhancing the by Van Putten et al. proposed slowing and asymmetry detection.

The proposed method to enhance SpikeNet is twofold. SpikeNet, a convolutional neural network, is firstly trained on 9005 control patients and 88297 candidate IED's. To reduce the false positive rate, hard example mining was applied. This is a method where you predict your training set with the freshly trained model, to identify wrongly predicted EEG segments. The wrongly predicted EEG segments are added to the dataset and the model is trained again. The latter steps are repeated 15 times, resulting in our final model called $SpikeNet_{15}$. A 70% false positive reduction is found resulting in a false positive rate of 15 per hour at a sensitivity of 95%. Using this overwhelming evidence, we conclude that hard example mining increases the model performance significantly and making this a crucial step in training similar models.

Secondly, we tried to enhance the performance of SpikeNet by adding generated EEG segments to the training set. We build three versions of Generative Neural Networks (GAN's); a GAN and a Wasserstein GAN with gradient penalty which are both optimized using ADAM, and finally we build a Wasserstein GAN with gradient penalty which is optimized using Adamod (WGANGP-Adamod). WGANGP-Adamod outperformed the other GAN versions and was able to increase the area under the ROC curve (AUCROC) from $SpikeNet_{15}$. Even though the AUCROC increased, the performance increased, the FP/h at 95% decreased from 15 to 18.3 FP/h, resulting in a decreased performance of $SpikeNet_{15}$.

The second contribution we made was the validation and enhancing of the *BSI* and *tBSI* which are, respectively, the asymmetry and slowing detection algorithms proposed by Van Putten et al. Both detection algorithms are relying on the power of the EEG of the patient. The *tBSI* requires a healthy reference EEG from the same patient before the slowing calculations can start. This dependency makes this algorithm useless if no reference EEG from the same patient is present. In order to overcome this burden, we generated a reference matrix using a neural network. The reference matrix includes the average power of the EEG, per frequency, per channel, per age. This reference matrix does include all the necessary features making a reference EEG useless.
After implementing the reference matrix in the *tBSI*, we optimized the bandwith for the power calculations in both the *BSI* and *tBSI*. The *BSI* and *tBSI* are both evaluated using their own dataset of 200 patients. Each dataset contained 100 control patients and 100 (near) continuous slowing/asymmetry patents. Evaluating the algorithms lead to an AUCROC of 0.95 for the *BSI* and 0.88 for the *tBSI.* Further research is needed to make these algorithms applicable for intermitted slowing.

Preface

After approximately eight years of studying at the University of Twente, I finally come to the point of defending my thesis and signing my diploma. I have to say, it went by so fast! Those eight years have been an incredible time for me, I have learned so much, had a small sidetrack into MII and visited Asia and America for university purposes, but I am back now. Temporarily.

In the past 11 months, I did my graduation internship at the Neurology research department of the Massachusetts General Hospital. It truly was an exciting experience in the broadest sense of the word. During my thesis I was able to develop myself in the field of Deep Learning and Interictal Epileptiform discharge detection and I want to thank a number people for that.

First of all, Brandon Westover, thank you for sharing your endless ideas and knowledge. You are a great motivational supervisor and I really appreciate your love for research and that you can be genuinely happy for someone if he or she made any progress. Jin Jing and Weilong Zheng, thank you for all the talks that we had in real life and via zoom. I could always come to you with questions regarding the Interictal epileptiform discharge detection and Generative Adverserial Networks.
I belive Brandon Westover, Jin Jing and Weilong Zheng, deserve a second thanks, for reviewing the thousands of false positive predictions that I send you, in such short notice. This human evaluation really elevated the project and this tremendous task was always completed within a few days. Haoqi Sun, thank you for answering my programming and computer related questions. I am still flabbergasted by thinking about the time that you fixed my problem before I could even ask you.
Bregje Hessink-Sweep, thank you for always being able to touch the raw nerve. You made me dig deeper into my behavior, enabling the discovery of new insights about myself. Michel van Putten, thank you for being my technical supervisor and the chairman of my graduation committee and last but not least, Elena Mocanu, thank you for being the external member of my graduation committee.

Table of content

Abbreviations

AUC	: Area Under the Curve
AUCROC	: Area Under the Receiver Operator Characteristic Curve
AUCPR	: Area Under the Precision Recall Curve
BSI	: Brain Symmetry Index
CAR	: Common Average Reference
D	: Discriminator
DB	: Double Banana
EEG	: Electroencephalogram
EM	: Earth Mover distance
FP/h	: False Positives per Hour
FP/m	: False Positives per Minute
DALY	: Disability-Adjusted Life-Years
G	: Generator
GAN	: Generative Adversarial Network
GAN-ADAM	: Generative Adversarial Network with ADAM as optimizer
Grad-CAM	: GRADient-weighted Class Activation Mapping
GUI	: Graphical User Interface
IED	: Interictal Epileptiform Discharge
FID	: Frechet Inception Distance
IS	: Inception score
POSTS	: Positive occipital sharp transients of sleep.
PR	: Precision Recall
PSD	: Power Spectral Density
ReLU	: Rectifying Linear Unit
ROC	: Receiver Operator Characteristic
r-BSI	: Revised Brain Symmetry Index
r-tBSI	: Revised Temporal Brain Symmetry Index
SGD	: Stochastic Gradient Decent
$SpikeNet_n$	: The n^{th} version of SpikeNet
tBSI	: Temporal Brain Symmetry Index
VAE	: Variational Auto Encoder
WGAN	: Wasserstein GAN
WGANGP-ADAM	: Wasserstein GAN with Gradient Penalty which uses ADAM as optimizer
WGANGP-Adamod	: Wasserstein GAN with Gradient Penalty which uses Adamod as optimizer

Thesis introduction

In 2016, neurological disorders where the second leading cause of global deaths with 9 million annual deaths, and where the leading cause in disability-adjusted life-years (DALY) with approximately 276 million DALY's. Over the course of 1990 to 2016, a 39% increase in deaths and 15% increase in DALY's is found. The prevalence of neurological disorders steeply increased with age, with an increasing world population and life expectancy a further increase in deaths and DALY's is imminent [1]. This will lead to an increased demand of the already scarce qualified personnel leading to the need of new prevention and treatment strategies [2].

One of the most commonly used techniques is the electroencephalogram (EEG), which is able to non-invasively measure brain activity [3] and is widely used for the diagnosis of, but not limited to, epilepsy[4]–[7], traumatic brain injury[8], [9], stroke [10], [11],encephalitis [12], [13],brain tumor[14], [15], encephalopathy [16], [17], memory problems [18], [19], sleep disorders[20]–[23] and coma [24], [25]. Visual inspection is still the golden standard for the clinical interpretation and analysis of the EEG [26] and during visual analysis of the EEG one must account for reactivity, symmetry, synchrony, morphology, the level of occurrence and localization of certain EEG patterns [27]. It is not hard to imagine that reading an EEG must be done precisely and Brogger et al. showed that reporting a routine EEG takes on average 12.5 minutes [26].

Visual scoring is subject to the interpretation of the expert resulting in different outputs of the same EEG while scored by different raters. This interrater variability differs drastically from task to task and Westhall et al. reported a kappa of 0.71 for determining highly malignant patters, 0.72 for rhythmic or periodic malignant patterns, 0.42 for malignant patterns and 0.26 for unreactive EEG [24]. Using automated EEG analysis techniques as stand-alone feature or as a supplementary one, will save time and will increase the output consistency. Also, knowing that in developed countries, the number neurologist per 100.000 inhabitants varies between 1 and 10, in major parts of the world, mostly Africa and South East Asia, neurology is marginally present [2]. Therefore, automatic analysis of the EEG will reduce the burden of the Neurologists in developed countries but also elevate neurology in the less developed world.

Fully automating the EEG analysis is a project too big to be handled on its own, so many studies focused on automating a subtask, for example diagnosing a single disease [4]–[15], [20], [21], [23].
This thesis is a contribution towards a fully automatic EEG analysis and is containing two subtasks; Automatic interictal epileptiform discharges (IED) detection (chapter 1 & 2) and generalized / localized slowing detection (chapter 3). Even though multiple projects are addressed, the main focus of this thesis is the IED detection.

IED detection is already addressed by multiple studies where Jing et al. does have the best results at the time of writing with a deep neural network called 'SpikeNet' [4], [6]. Even though the results of Jing et al. do surpass the expert level performance, the false positive rate does leave us some room for improvement [4].

Improving an already existing model can be done in three ways: Alternating the model, alternating the data or alternating both the model and the data. The excellent performance of SpikeNet does suggest a well-chosen architecture witch sufficient usage of the data. If the architecture and original data is fully exploited, data augmentation has proven to be an effective way to enhance the performance of already existing models [28]–[31].

Data augmentation is an umbrella term for alternating your data into different, useful data. This augmentation can vary in complexity and ranges from rotating and scaling data, up to synthesizing new data [28]. Most data augmentation techniques are solely used for enriching one or more classes in the dataset with the main goal of increasing the generalization and/or countering the class imbalance [32]. If a class includes a broad range of patterns that should lead to the classification of that specific class, such as the multiple morphologies in IED detection, it is found that not all patterns are equally hard to classify correctly. The patterns that are harder to classify, or so-called hard examples, do potentially yield important predictive values. Localizing and adding these samples to the dataset will gradually increase its difficulty, which may lead to an increased performance [33].
The amount new data that can be created using hard example mining is limited, since technically no new data is created, which motivates the choice for additional technique.

A less limited augmentation method is the use of Generative models. Generative models, as the name suggest, are able to generate data by learning the distribution of the data [34]. In this way generative models are able to generate new plausible data. In recent years, the interest in generative models has drastically increased as result of the state-of-the-art performance that they deliver [35], [36]. By combining both the hard example mining and generative model, it is possible to gradually increase the difficulty of the dataset and enriching one or more classes. Both techniques have proven to be effective in other fields, however they are not applied for this specific application yet [30], [35], [37].

Research question

Based on the previous, we could state the following research question:

'To what extent will advanced augmentation methods contribute to the improvement of an already state-of-the-art interictal epileptiform discharge detector?'

We can further divide this research question into the following sub questions that each will be discussed in separate chapters.

- *To what extend will using a semi-automatic hard example mining method, reduce the false positive prediction of the interictal epileptiform discharge detector? ~ Chapter 1*

- *To what extend can generated EEG segments increase the performance of the interictal epileptiform discharge detector? ~ Chapter 2*

Enhancement of the interictal epileptiform discharge detector via hard example mining

Introduction

Over the years, EEG has established itself as an essential non-invasive neuronal diagnostic tool. It is mostly used to diagnose epilepsy but can help determine sleep disorders, depth of anesthesia, coma, encephalopathies, and brain death [38]. When epilepsy is suspected, an EEG measurement is recorded, and a certified physician will look for abnormal EEG patterns with an epileptic nature. Abnormal EEG patterns can present in many forms, however IED's are a typical display of an abnormal EEG with an epileptic nature [39]–[41]. IED's include, but are not limited to: spikes, sharp waves, benign epileptiform discharges of childhood, spike–wave complexes, polyspikes, hypsarrhythmia and seizure patterns. Usually, abnormal EEG's with an epileptic nature are found in 50% - 88% of patients with epilepsy during a single EEG measurement; repeated EEG's, long time recordings and activation procedures will increase the chance of recording IED's [42].

For most of the EEG analysis, including the identification of IED's, manual analysis by a specialized physician is still the golden standard. Manual scoring is a time-consuming activity and is subject to inter- and intra-rater variability [5], [43], [44]. It is also seen that manual classification is a dying phenomenon in the era of computers, where many processes are getting automated, including the analysis of medical data [4], [20], [21], [23], [45]. In the last decade, multiple computer-based models are developed to automatically analyze EEG recordings, covering a wide range of applications among which IED detection algorithms present [4], [20], [21], [46]. Jing et al. developed SpikeNet, an IED detection algorithm which results surpassed both the expert interpretation and industry standard [4].

A major problem for the current automatic IED detectors is the false positive rate, making these automated IED detections unsuitable for stand-alone clinical usage [6], [47]. A similar problem is well known in the clinical setting considering that distinguishing between true IED's and benign variants of uncertain significance is perhaps the most challenging for novice physicians [48], [49]. Indicating subtle morphological differences between the IED and benign variants of uncertain significance.

There are many ways to elevate the performance of your model. However, the most promising method to reduce false positives is hard example mining [33], [50], or in other words, gradually improving the difficulty of the dataset by adding incorrect classified samples. This approach, firstly described by Sung et al. [51], will require an unknown number of training iterations and will be finished when convergence or a performance drop is reached [33], [50]. In this work, we retrospectively evaluate, using the MGH clinical care dataset, to what extend using a hard example mining method, will reduce the false positive prediction.

Methods

Dataset

Retrospective analysis of the EEG data was approved by the Partners Institutional Review Board without requiring additional consent for its use in this study. The data was recorded as part of routine clinical care in the MGH neurology department from 2012 until 2018. All EEG's in the presented analysis are recorded using equipment from Grass Technologies (now owned by Natus Neuro, CA, US). EEG electrodes were placed in the following 19 locations according to the international 10-20 system: Fp1, F3, C3, P3, F7, T3, T5, O1, Fz, Cz, Pz, Fp2, F4, C4, P4, F8, T4, T6 and O2. Each EEG was reviewed by an EEG laborant and/or a physician. Each patient was labeled as spike/non-spike and normal/abnormal, respectively according the presence of interictal spikes and the presence of abnormalities in general.

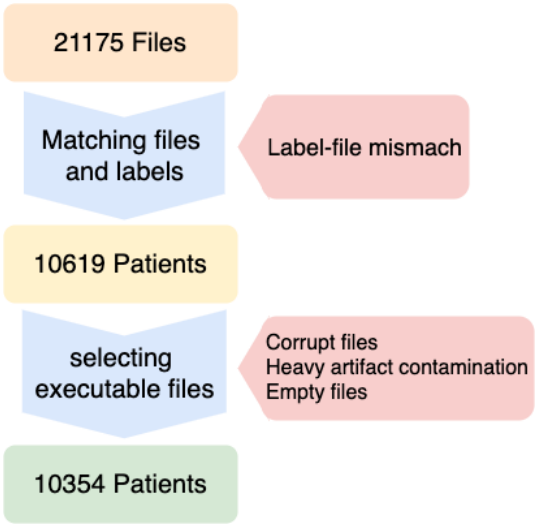

Figure 1. 1 Schematic overview of patient inclusion

The MGH EEG dataset contains 21175 patient files, where long measurements are cut into multiple files. After matching the files with the available labels, 10619 patients were selected. 10370 files where pre-processable. After removing heavily artifact contaminated EEG's, a total of 10354 EEG's are used in this study. A schematical representation of the inclusion pathway is given in figure 1.1.

Pre-processing

Signal pre-processing

The raw EEG is resampled to 128Hz, after which a high pass, low pass and a notch filter are applied of 0.5Hz, 64Hz and 60Hz respectively. After filtering, the Common Average Reference (CAR) montage is applied and the EEG is clipped between -500mV and 500mV.

Dataset pre-processing of the control group

Patients without IED's are selected for the deep learning dataset. We randomly sampled 2000 non-IED examples of 1 second per patient. If 2000 samples exceed the number of samples in the corresponding EEG, the maximum number of samples is taken. These non-IED data will be referred to as control data later on.

Dataset pre-processing of the IED's

The neurology research department at the MGH has a database of 88297 candidate IED's.
The 88297 candidate IED's can be reduced to 13262 morphologically distinguishable candidate IED's. The 13262 morphologically distinct candidate IED's are rated by 8 experts, where each expert will score the candidate IED's as 'IED' or 'No IED'. Combining these scores will result in a soft label between 0/8 and 8/8 where the fraction stands for the number of raters that scored the candidate IED as an actual IED.
The 13262 labeled candidate IED's will be referred to as medoid spikes. The 75035 candidate IED's that are not labeled, can be clustered and linked to a medoid spike using the morphological similarity, resulting in 13262 clusters with one medoid spikes and multiple member spikes. Each member spike will receive the same label as the medoid in their cluster to ensure all 88297 candidate IED's are labeled.

To enlarge the dataset and increase the variety of the candidate IED's, the medoid and member spikes are augmented. First, the left and the right channels are switched in the montage, secondly the waveform was translated $\pm$ 0.1 second in time.

After labeling, the data is split, patient wise, in a train, test and validation set.
Each set contains respectively 70%, 15% and 15% of the control and spike patients. To ensure the performance evaluation of the model is not disturbed by augmentation or weak labeling, only control and non-augmented medoids are used in the validation and test set. The training set uses the control data as well as the augmented and non-augmented medoids and member spikes.

Network architecture

SpikeNet, the convolutional neural network created by Jing et al. is used in this study. The architecture of SpikeNet is based on Hannun et al. [45]. The input of the model is a one second EEG segment sampled at 128 Hz containing 19 CAR channels and 18 bipolar montage channels. A single dimension is added to gain a three dimensional matrix with the size of [128,1,37]. This extra dimension is necessary when applying two dimensional convolutions and is used accordingly.

The first block finishes with two consecutive convolutional layers. The first convolutional layer carries out a temporal convolution whereas the second layer carries out a spatial convolution. This repeated convolutional layer, which is based on Schirrmeister et al. [52], is applied so that all channels have the same temporal kernel. The separation of temporal and spatial convolutions may increase performance for EEG signals [52].

The second block, which is used twice, is also known as a residual block. Each residual block increases the number of filters by 32 and reduces the time dimension by a factor of 4.
The batch normalization centers and scales the data within a batch to a zero mean and a unit standard deviation.
After batch normalization, a leaky rectified linear unit (ReLU) is applied as an activation layer. To improve regularization, a dropout layer is implemented that ignores 20% of the incoming nodes

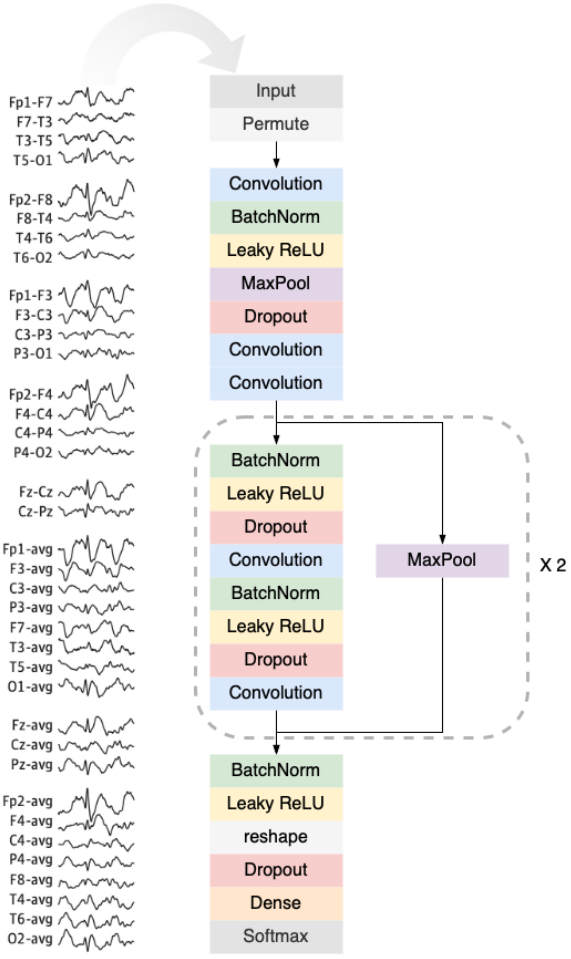

Figure 1. 2 Model architecture of SpikeNet

In the last block, the data, consisting the output of the last convolutional layer, is prepared for classification. This is achieved by reshaping the three-dimensional data into a one-dimensional array so it could fit into a dense layer. After the dense layer a SoftMax layer is applied. The SoftMax layer outputs the predicted label $\hat{y}$ which is the calculated chance if a spike is present in the input data.

Validating the model

Validation of the model the sensitivity, calculated only using medoids, is plotted against the false positive rate per minute, calculated on the control data. This curve will be referred to as ROC_{adjust}. An adjusted PR curve, later referred to as PR_{adjust}, is calculated only using medoids for the true positives & false negatives, and only control data for the false positives.

Processes of the Iterative training

The training is an iterative process where each iteration consists of multiple actions. All possible actions that are used, are described in this paragraph. However, small differences between training iterations are present and a full overview of the training iteration is given in figure 1.5.

Train: Train the model using the training and validation set.

Predict: The new model predicts the training, validation and test set.

Background rejection: An inhouse build, rule-based background rejection algorithm is applied on the predicted output to filter out artifacts. More in-depth information of the background rejection is given in appendix 1.

Visualizing the convolutional focus: In the first round, Gradient-weighted Class Activation Mapping (Grad-CAM) [53] is used to highlight the EEG segments that were important for the prediction. This technique visualizes the convolutional focus and shows us if the model focusses on the correct parts of the EEG. Additionally, it give us insight in the patterns that result in false positive predictions. More in-depth information for Grad-CAM is given in appendix 2.

Automatic performance validation: For the automatic performance validation, the ROC_{adjust} and the PR_{adjust} are calculated. Both graphs are calculated multiple times using a different range of IED's. The graphs are calculated using IED's with the label ranges, ranging from IED's with the label 5/8 and higher, to only IED's with the label 8/8.

Manual performance validation & label enhancement: In the first 4 and last round, manual performance validation is applied complementary to the automatic performance validation. For the manual performance validation all control patients are used. Manual inspection is applied to see what kind of patterns causes false positive predictions. It could also lead to the finding of incorrect labeled patients. To get the insight in the false positives and to have the opportunity to relabel patients, a graphical user interface (GUI) was built in MATLAB. For each patient, as many false positives as possible with a maximum of 3 are manually inspected. The GUI and the visualization of the EEG are shown in figure 1.3 and 1.4.

Figure 1. 3 The GUI that was build and used to label candidate false positive segements

Relabeling: After the manual performance validation a relabeling step is performed. Patients with IED's present in their EEG are relabeled in the database and are removed from the control dataset.
Enhancing the dataset: Subsequently to the relabeling, the hard examples of the spikes with the label "No-IED" are added to the dataset. All patients that are relabeled from "No-IED" to "IED" are removed from the dataset.

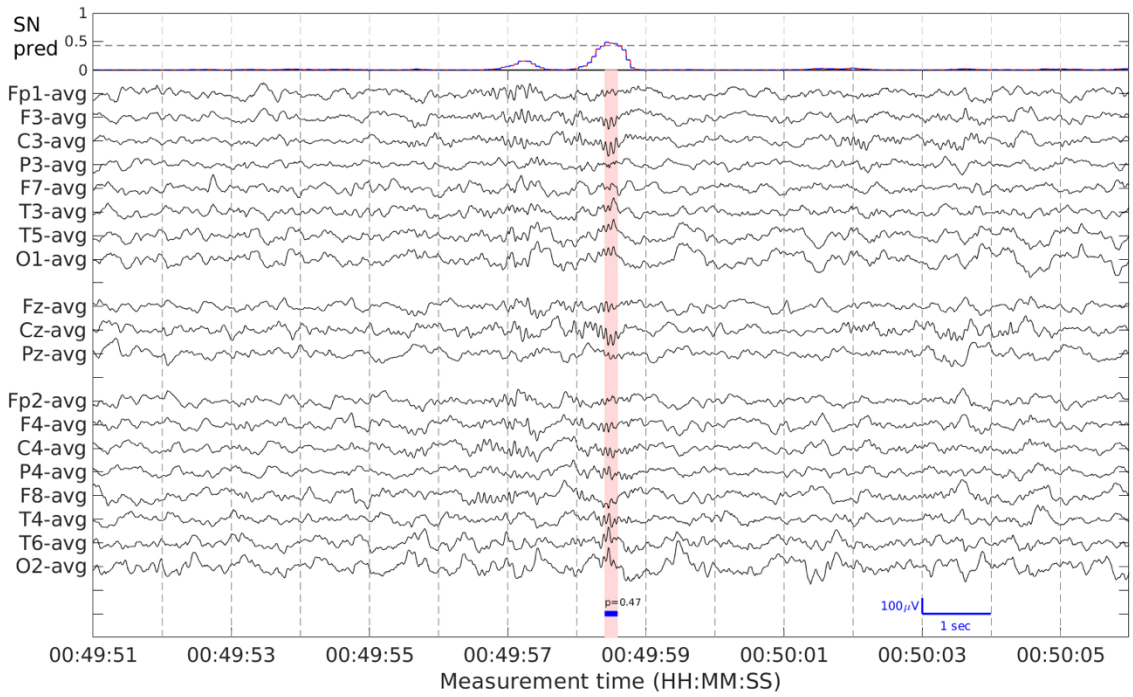

Figure 1. 4 The candidate false positive segment as shown to the raters

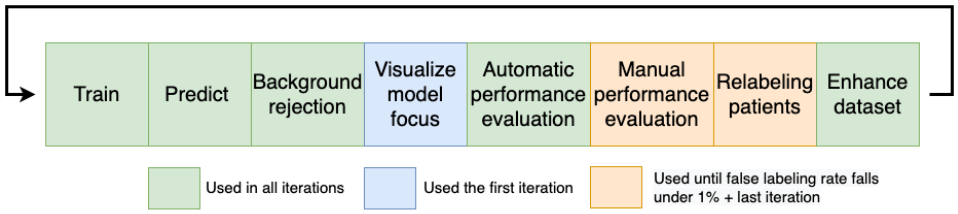

Figure 1. 5 A schematic overview of which elements are used during the training iterations

Training the model

The enhancement of the model is an iterative process, the training is stopped if no positive performance trend is present over the last 3 iterations based on the automatic performance. In the first training iteration, 6285 control patients are used, containing a total of 4.806.215 EEG segment. Manual inspection was applied during the iterations to find wrongly labeled patients. Manual inspection is applied until the false labeling rate was under 1% of the total number of patients. Since relabeling patients and adding hard examples effects the size of the dataset, an overview of the used data per iteration is given in table 1.1.

Results

Patient demographics

During the training, the demographics of the dataset changes due to the relabeling. Control patients are excluded from the dataset of IED's was found within their EEG. As seen in table 1.1, a total of 753 patients are relabeled and therefore excluded as control patient.

Table 1. 1 Demographics of the dataset.

IED's	Total number of medoid candidate IED's	13262
	Total number of member candidate IED's	75035
Control patients	Number of control EEG's used at the start of iteration 1	9005
	Number of control EEG's used at the start of iteration 2	8475
	Number of control EEG's used at the start of iteration 3	8305
	Number of control EEG's used at the start of iteration 4-15	8252
	Mean measurement length ± std of the 8252 control EEG's (min)	57.5 ± 24.1
	Mean age ± std of the 8252 control EEG's (age)	43.1 ± 26.7

Visualizing model focus

In the first training iteration no hard examples are present. During this iteration, we evaluated the performance of SpikeNet which will be used as baseline performance later on. As a sanity check, we visualized the convolutional focus to better understand the model predictions. The convolutional focus is plotted as a heatmap over the corresponding EEG. If the focus of the model increases, the heatmap will converge from blue via yellow to red respectively meaning low, moderate and high focus. Segments with labels ranging from 0 (0/8 & control) to 1 (8/8) are evaluated, only one per label is visualized below. For each segment, the label and the predicted value are given above the EEG. Looking at the convolutional focus, it can be seen that the candidate IED's are highlighted in red from the label 3/8 and above. Candidate IED's with labels below 3/8, are highlighted in yellow or not highlighted at all. Looking more closely at the convolutional focus, it is found that the model mainly focuses on a sudden upward transition. The convolutional focus is increased if the upward transitions are simultaneous in multiple channels. If the transition is different and/or not present in multiple channels, mostly yellow representations are found meaning moderate focus of the model.

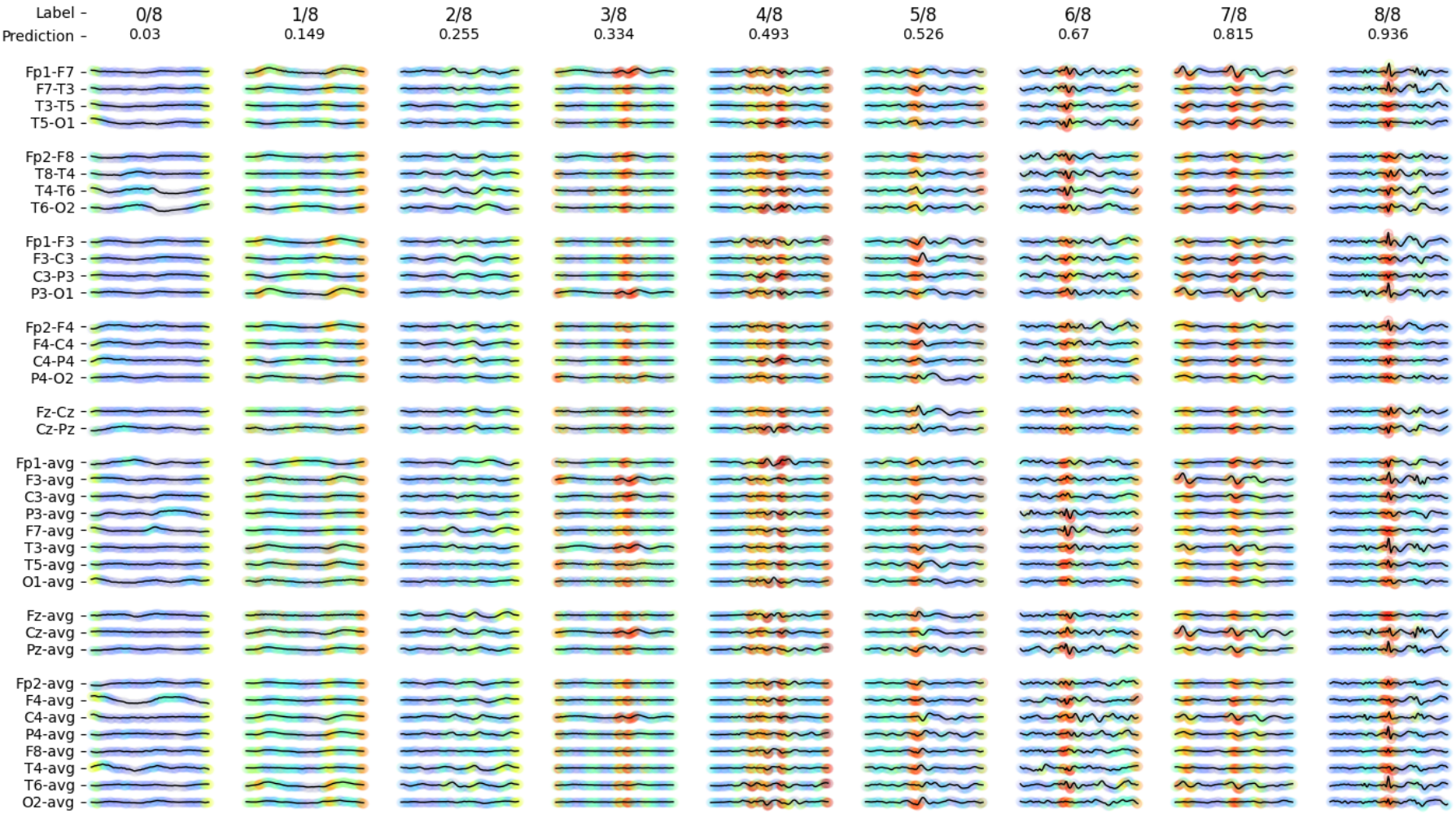

Figure 1. 6 A visual representation of the model focus. 9 segments with labels ranging from 0/8 until 8/8 are included. For each segment the prediction is plotted and the model focus is shown using an heatmap overlay. Where blue means low focus and red means high focus. As seen in the model, the high focus, in red, is mostly present at sudden upward transitions.

Manual Performance evaluation

During the manual performance evaluation, a subgroup of all false positives is subjected to visual inspection. During this inspection, multiple recurring patterns that will produce high model predictions were found. All recurring patterns will be separately discussed below.

Artifacts that effect (almost) all leads during the measurement

Artifacts are commonly present in EEG's and cause the majority of the false detections in SpikeNet's prediction. Artifacts which affects all leads are regularly found in EEG's and are most likely to a movement artifact. In general, high voltage movement artifacts are captured by our background rejection model and the model does in general not return high prediction when confronted, however not all artifacts are captured and, in some cases, high predictions are returned.

The upper panel of figure 1.7 shows the model prediction where the red curve is the predicted output of SpikeNet, and the dotted black line is the predicted output after background rejection. The dotted black line and the red line will overlap if the data is not rejected by the background rejection algorithm. As seen, the high voltage artifacts before 0:06:58 are captured by the background rejection and the model does predict values up to 0.5. When the high voltage EEG artifacts reduces, the background rejection fails to reject the prediction and the SpikeNet prediction crosses the threshold value of 0.43 as indicated in the middle of the EEG by the vertical red band.

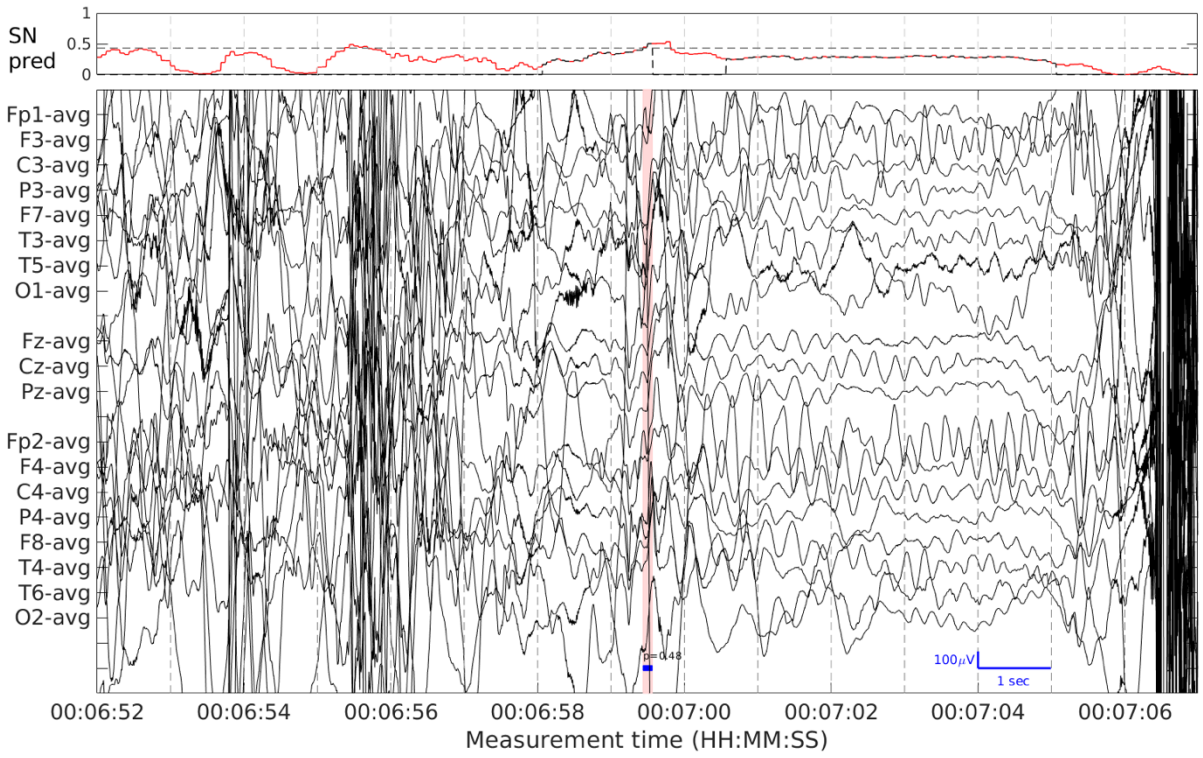

Figure 1. 7 This false prediction is caused by artifacts. In the upper panel, the red line shows the model prediction and the blue dotted line shows the model prediction after background rejection. it can be seen that the background rejection fails to reject all artifacts resulting in a threshold crossing prediction at 00:06:59.

Artifacts during the measurement which effect only a small number of leads

Sharp and transient artifacts which only effect a small number of leads can create spike like behavior when the CAR montage is used. Even though the model receives 2 montages CAR and double banana (DB) montage, the model is fooled by this artifact. The enormous spike present in T4, reaching up to Fz in the montage, creates a brief very high average resulting in the downwards spikes in all other channels, as can be seen in figure 1.8. The intensity of the artifact and the number of affected channels are key factors for the model prediction. When the artifacts effect only one ore a few electrodes for a brief moment of time, SpikeNet predicts much higher values (0.7), compared to a higher voltage artifact that indicate more channels as shown in figure 1.7.

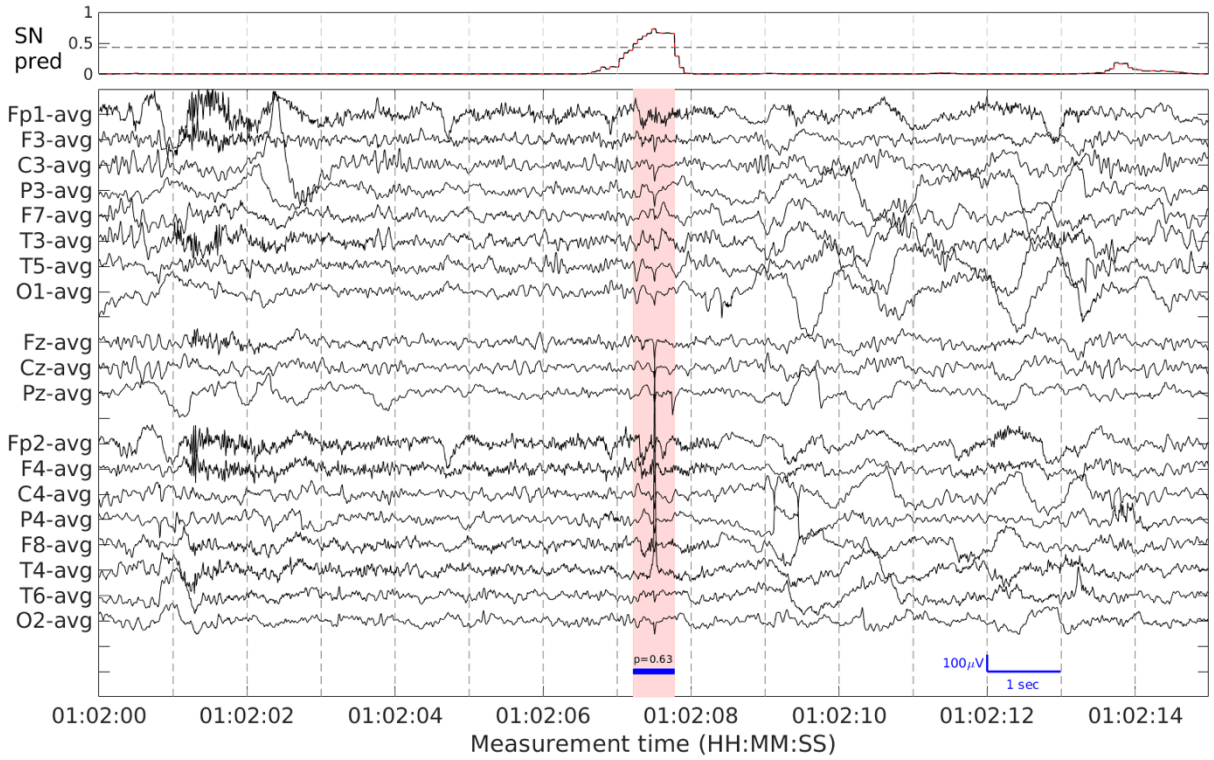

Figure 1. 8 A artifact present in the lead T4 will create a spike like pattern using the common average reference montage as seen at 01:02:07.

Artifacts that are induced by starting and ending the measurement

The MGH EEG dataset pre-processing steps, as described earlier, do not include clipping the EEG at the start and end of the measurement. Resulting in a series of artifacts that are not present during the measurement itself but, non the less, are present in our false positive evaluation. Artifacts created by starting or ending the measurement are characterized by an abrupt start of end of the EEG.

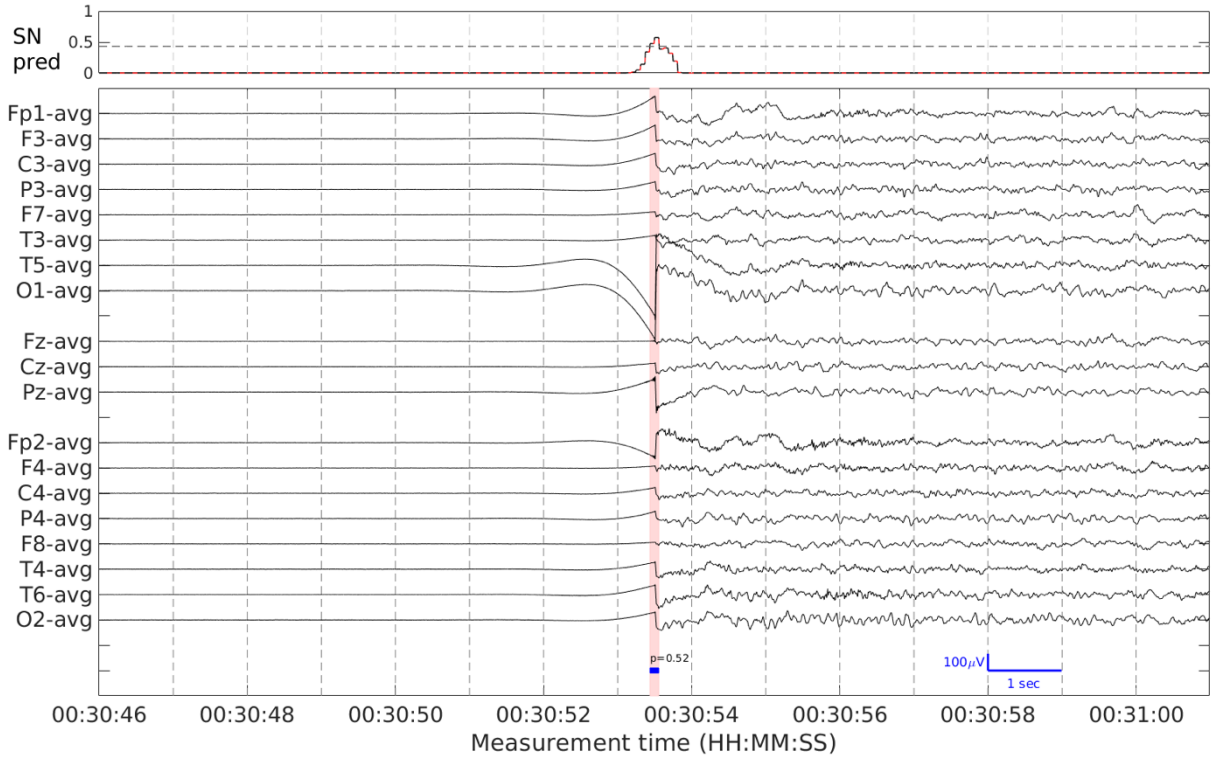

Figure 1. 9 When the EEG recording starts, a typical starting artifact is created. Similarly, when the EEG recording ends, a mirrored version of this artifact is present. Due to the sharp transition, the artifact is falsely detected as a IED

Artifacts that are induced by the calibration of the EEG

When the EEG is turned on, but before the measurement starts, the EEG equipment needs to be calibrated. The mechanical calibration of the EEG signal leads to a sinusoidal waveform in the prediction due to the consistent changing EEG.

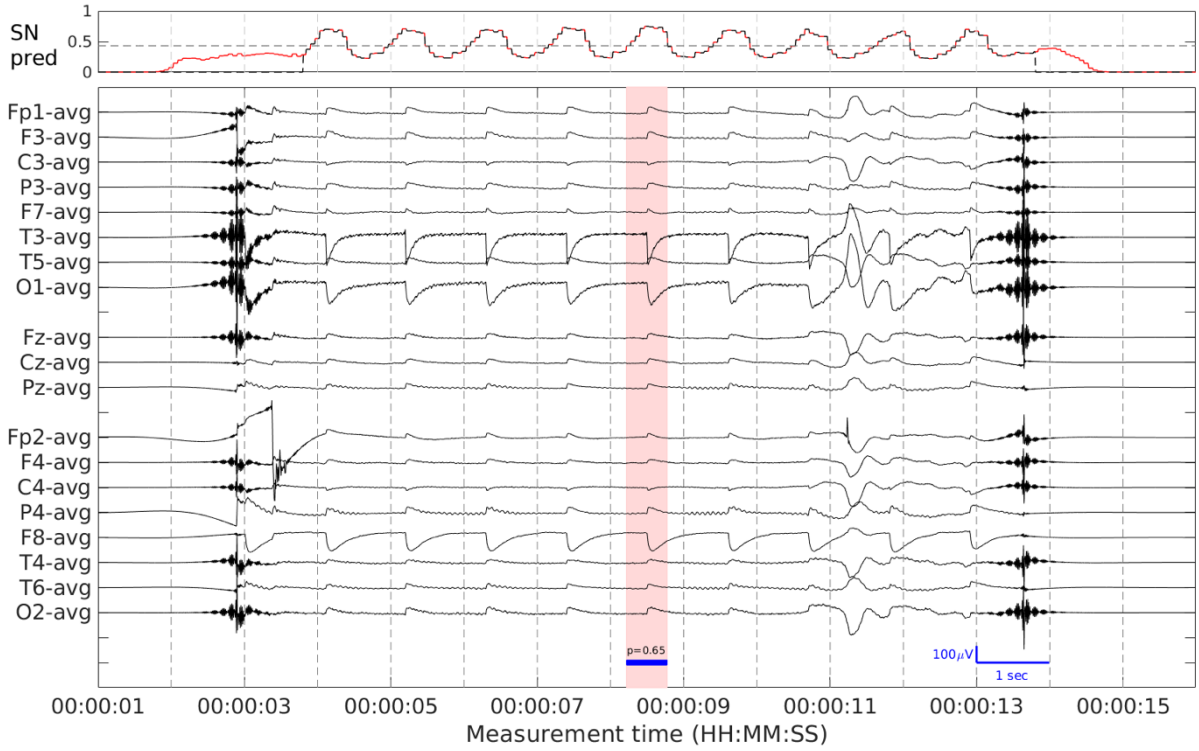

Figure 1. 10 A similar morphology compared to the artifact shown in figure 1.9 is created by the calibration of the EEG which also lead to a false detection.

False positives caused by benign variants of uncertain significance

Some EEG patterns might be epileptiform appearing, since their morphology appears to be a sharp waveform or a spike. However, these patterns do not yield any relationship to epilepsy. The appearance of these benign variants of uncertain significance will become clinical significant if they are over interpreted and mistaken for IED's [54]. Benign variants of uncertain significance which repeatedly causes false positive predictions, are described in below.

Hypnagogic hypersynchrony

Hypnagogic hypersynchrony is a hallmark of drowsiness in children aged 3 to 13 years. It can be described as generalized, paroxysmal, synchronized, high voltage, slow wave activity which lasts around 2 to 8 seconds.[55] During the hypnagogic hypersynchronisation, the slow wave activity is synchronized to such an extent that the upward transitions occur nearly simultaneously. In addition to the synchronized upward transitions, the morphology also comes with a higher voltage than the background rhythms, creating steeper transitions, which is enough to trick SpikeNet into predicts values that are reaching up to the threshold value an above.

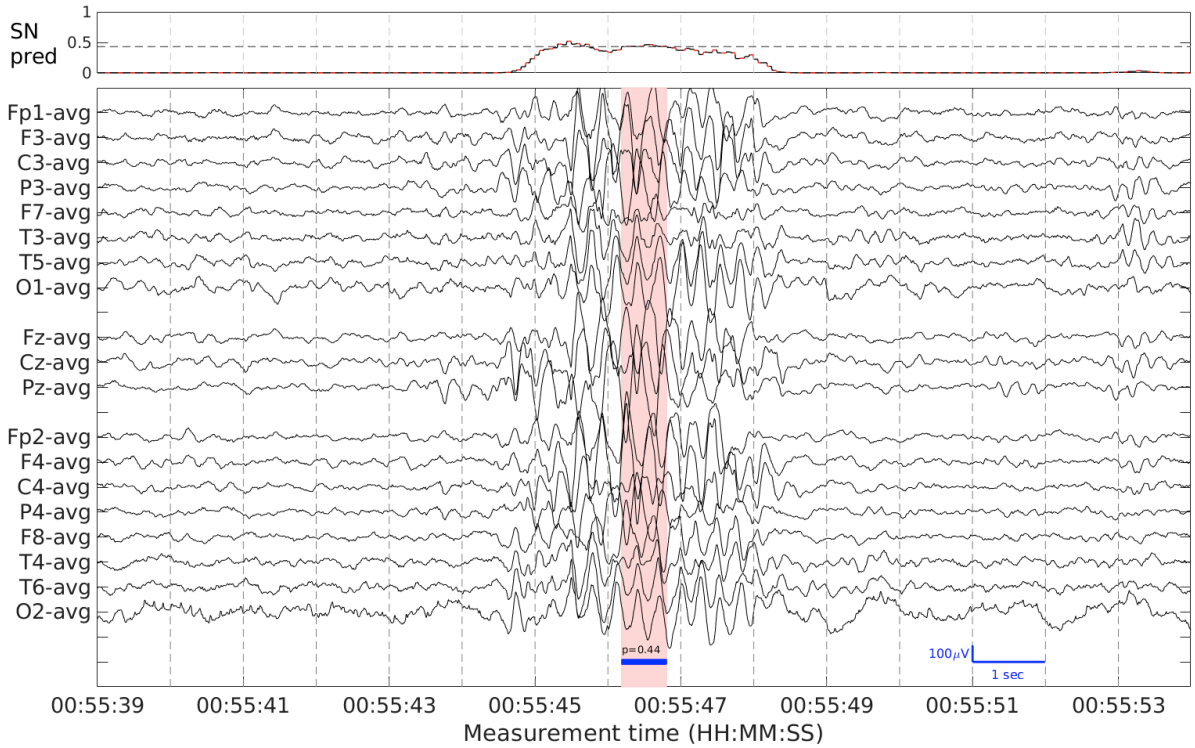

Figure 1. 11 The Hypnagogic hypersynchrony is clearly distinguishable between 00:55:44 and 00:55:49. The highly synchronized waveforms do tend to have the upward transitions at the same time, tricking SpikeNet into high output values.

Sleep spindles

Sleep spindles arise from thalamocortical oscillations and are a defining characteristics of stage N2 sleep. They have a frequency ranging between 11-16 Hz and lasting around 0.5 to 1.5 seconds. Drug spindles have a very similar morphology to sleep spindles, but slightly faster in frequency, and can be seen when benzodiazepines are administered [56]. Figure 1.12 shows an EEG in the sleep state with the presence of sleep spindles. The sleep spindle has sharp contours and does stand out from the background rhythm. When the amplitude variance of the spindle increases, SpikeNet is more likely to output higher prediction values.

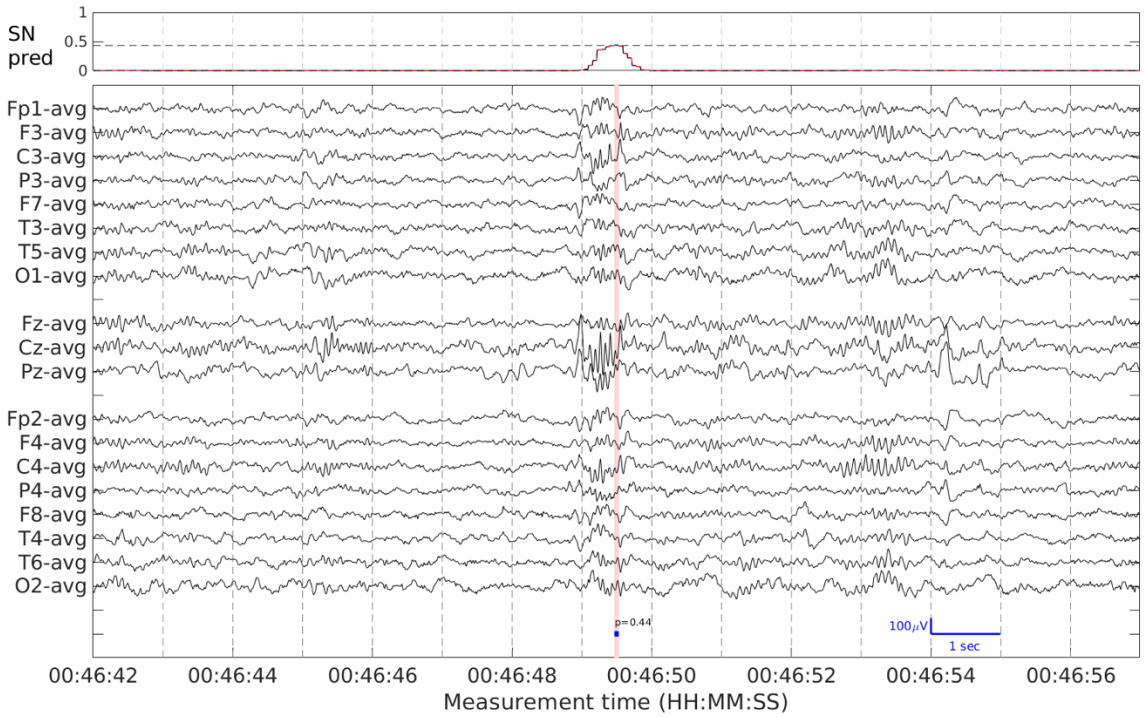

Figure 1. 12 A sleep spindle is present at 00:46:49. The sharp contours and steep slopes do mimic features of an IED to some extent. SpikeNet is tricked into predicting high output value by the 'spiky' appearance of the sleep spindle.

Vertex waves

Vertex waves are Sharply contoured waves finding their maximum over the central region of the brain and occur in late drowsiness and to some extend in N2 sleep [56], [57]. With a maximal duration of 0.5 second and a spiky appearance, they might mimic IED's in asymptomatic patients leading to incorrect predictions by SpikeNet. This is especially true in children due to the more spikey appearance of the vertex waves at younger age [54].

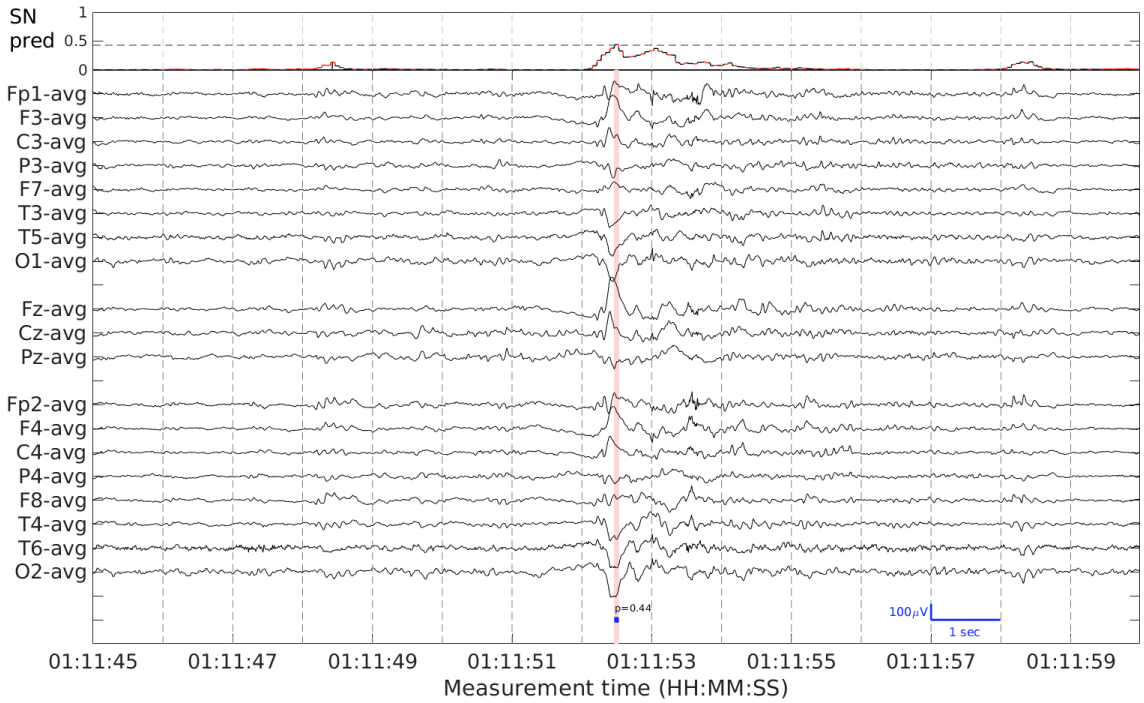

Figure 1. 13 The vertex waves, present at 01:11:52, finds its maximum amplitude at Fz-avg. The spikey behavior does lead to high, threshold crossing, output predictions.

Wicket spikes

Wicket spikes, mainly found during N1 and N2 sleep, are commonly present in trains with increasing amplitude of arciform waves with a frequency between 6 to 11 Hz [58]. They can also occur as a single waves, differentiation between an isolated wicket spike and an IED can be difficult due to similarities in the morphology which may lead to incorrect interpretation. [27], [54], [58].

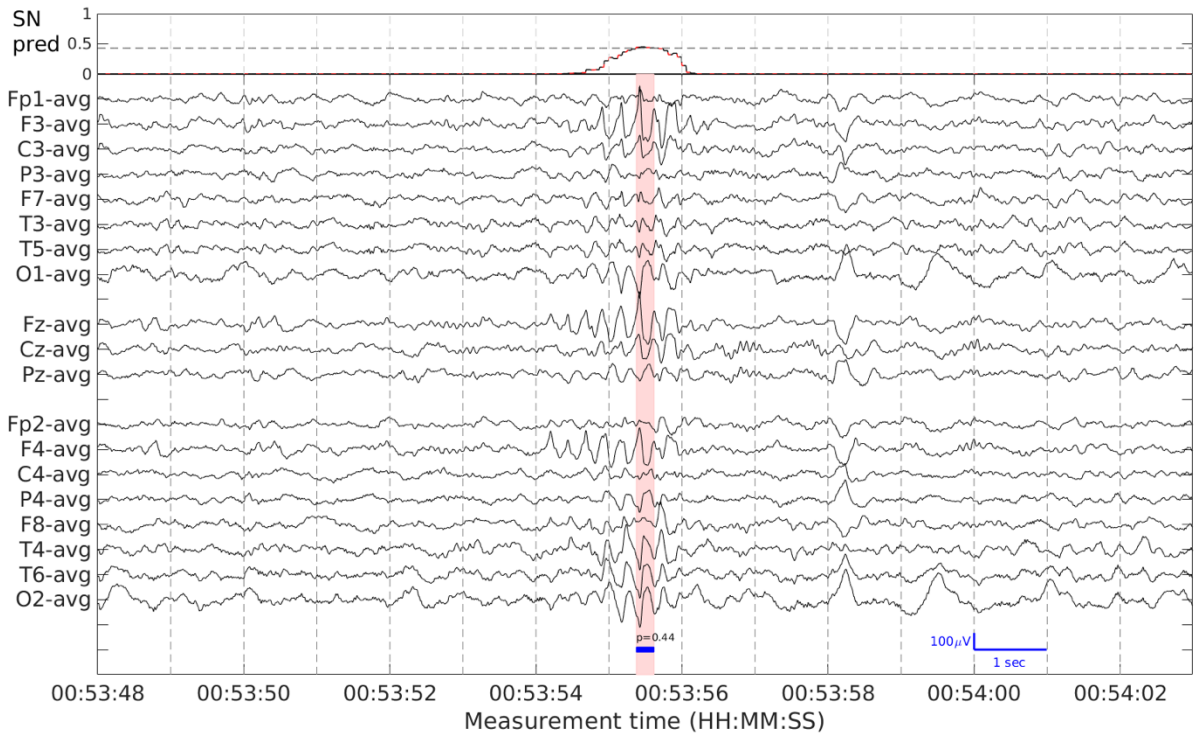

Figure 1. 14 A train of wicket spikes, present between 00:53:54 and 00:53:56, is clearly distinguishable from the background rhythm. The increasing amplitude is most noticeable in F3-avg and Fz-avg. It can be seen that the SpikeNet prediction rises as the amplitude of the wicket spike trains increases.

Automatic Performance evaluation

In the end of each training iteration, automatic performance evaluation is carried out using the ROC_{adjust} and PR_{adjust}. For each iteration, the area under the curve (AUC) for the ROC_{adjust} and PR_{adjust} is calculated using 1000 rounds of patient wise bootstrapping. Multiple ranges in the IED labels are considered in the calculations, to give a more complete overview of the model performance. The AUC of the ROC_{adjust} and PR_{adjust}, accompanied with their 95% confidence interval, are plotted against the training iterations to visualize the performance change as the training iterations increases.

As seen in both the ROC_{adjust} and PR_{adjust}, the third iteration yields a considerable performance drawback. Subsequently, an increasing performance trend is present in the following 12 epochs, overcoming the drawback and outperforming all previous models. The increased performance of the ROC_{adjust} and PR_{adjust} can be related to the decrease in false positives per hour (FP/h).

Figure 1.16 shows the FP/h per iteration at a 99%, 98% and 95% sensitivity level calculated using the same 1000 patient wise bootstrap as described above. Decreasing numbers of FP/h are found in all calculations among several levels of sensitivity. The greatest absolute reduction, of 42 FP/h, is found at a sensitivity of 99% calculated using only candidate IED's with the label 8/8. The greatest relative reduction, of 70%, is found using the candidate IED's with the label range of 5/8 – 8/8 calculated at a 95% sensitivity. A two sampled t-test is applied to evaluate if there is a statistical difference between the performance of 15^{th} and 1^{st} iteration of SpikeNet. We compared the bootstrapped ROC_{adjust}, PR_{adjust}, FP/h at 8/8 upto FP/h at 5/8-8/8 of the 1^{st} and 15^{th} iteration, all calculations where statistical different with $p<0.01$.

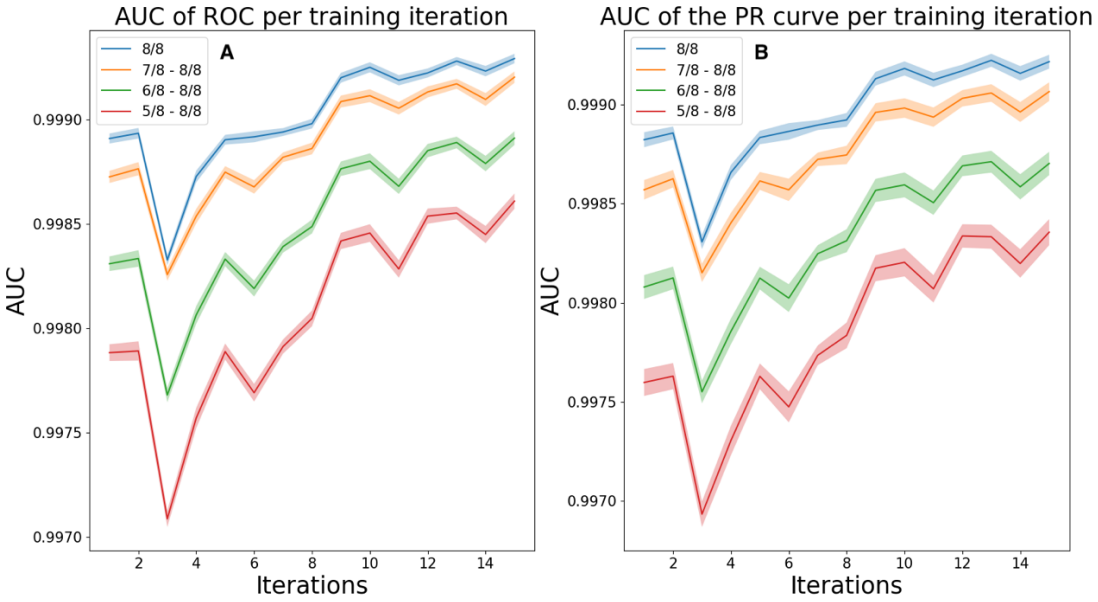

Figure 1. 15 The AUC of both the ROC_{adjust} and PR_{adjust} curves per training iteration. The bands that surrounds the line visualizes the confidence interval at 95%. The blue, yellow, green and red line are representing the AUCROC/AURPRC calculated using a different range candidate IED's. The candidate IED ranges are respectively, 8/8, 7/8 and 8/8, 6/7 to 8/8 and 5/8 to 8/8. The wider the range of candidate IED, the lower the AUC, which is expected due to the inclusion of less prominent IED's.

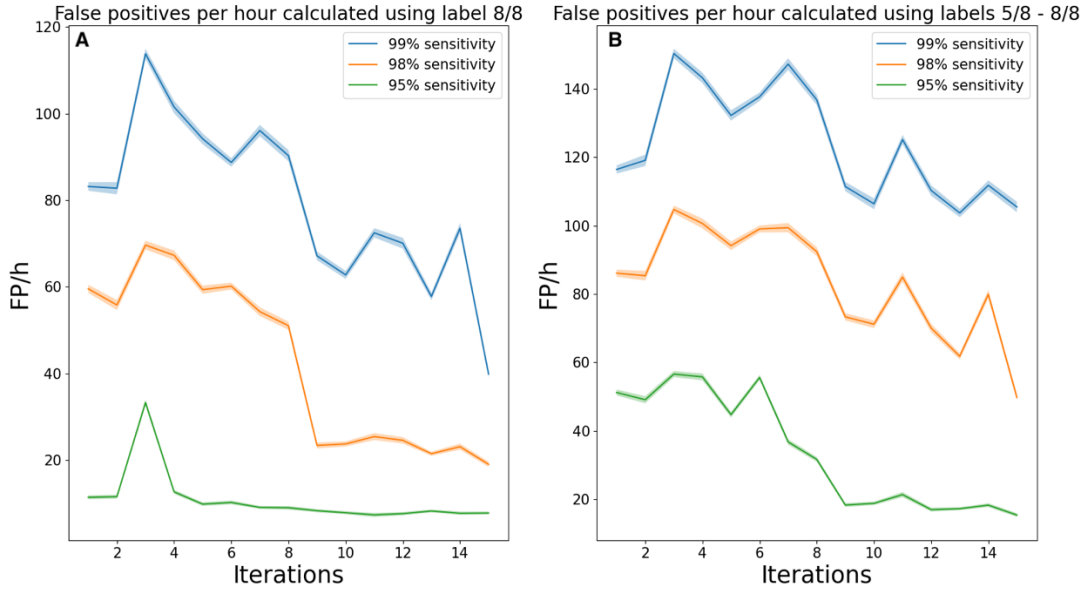

Figure 1. 16 The FP/h per iteration shown at 99%, 98% and 95% sensitivity. The bands that surrounds the line visualizes the confidence interval at 95%. The blue, yellow and green line are representing the FP/h calculated using at different sensitivities. sensitivities are respectively, 99%, 98% and 95%. As expected, the FP/h reduces as the sensitivity decreases.

Discussion

In our study, we enhanced SpikeNet resulting in an $AUCROC_{ajust}$ and $AUCPRC_{ajust}$ of at least 0.9985 and 0.9983 respectively. After 15 iterations of retraining, we succeeded in increasing the AUC of both the ROC_{adjust} and PR_{adjust} as well as decreasing the false positive predictions, resulting up to a 70% drop in false positive predictions per hour. Indicating that hard example mining and adding the mined examples during re-training is a successful strategy for increasing model performance without the need of acquiring new data. Suggesting that the proportion in the training data, corresponding to a data with a difficult level of distinguishability between 'IED' or 'No IED', and a possible increasing data diversity do play a critical role in the model enhancement.

Compared to earlier preformed studies, we did not limit our iterations to a pre-detained number. We seek to find the maximum number of training iterations for which a positive is present in the performance evaluation, and therefor confirming and extending the training method of Jing et al. Our model performance surpasses the performance of Jing et al. and therefor also the performance of experts using a partly similar dataset [4]. Our model, has a sensitivity of 95% ($95_{5/8}$%), calculated using candidate IED's with a label 5/8 or higher, while having a false positive rate of 15 FP/h.
Tjepkema-Cloostermans et al.[6] who is using a CNN-LSTM architecture. reports a 36 FP/h (0.6 FP/m) at a 47.4% sensitivity. Scheuer et al. reports that the Persyst P13, which is the golden standard for automatic IED detection, has a 43.9% sensitivity at 99 FP/h [47] and Hao et al. repots a 30 FP/h at a sensitivity of 84.2% while using EEG and fMRI. Our model is outperforming all well performing automatic IED detection known to us at the time of writing and therefor setting a new standard for automatic IED detection.

The strength of this research lies within the many iterations which allow tSpikeNet to carefully adapt to our enhanced dataset and increases its performance. In this training method, we created a harder training set every iteration by increasing the difficult examples in a semi-supervised way. This training method truly excels when all the false positives are checked by hand to make sure only true false positives are added. Since this manual validation takes a lot of time, a hybrid version, where 4 rounds are manually validated, is applied to make sure most test patients with actual IED's are excluded and relabeled for later studies.

A limitation our study is the lack of calibrating the IED threshold value during the iterations. If the optimal threshold has increased during the iterations, we have not included all false positives in our iterations, which may lead to a slower learning curve. On the contrary, if the optimal threshold for the IED detection has decreased over the iterations, we have falsely added true positives as false positives to our dataset.

During most training iterations, the model performance increased leading to a positive trend in performance. During the last round of manual validation, it appeared that the number of false positives created by artifacts is reduced more than false positives created by benign variants of uncertain significance. This can be explained when looked at the morphology of the false positives. The morphology of, for example, isolated wicked spikes, vertex waves or positive occipital sharp transients of sleep (POSTS) are more similar to an actual IED than to an artifact. Most artifacts can be easily spotted by shortly trained eye, however the benign variants listed above do tent do fool even the eye of experts [27], [54], [58]. The model can be seen as a new EEG expert in training, therefor it will first learn easy to learn discrimination features and later on, more sophisticated and fine-grained features will be learned leading to a better performance.

The problem with state-of-the-art automatic IED detection still remains the high false positive rates. Our method does reduce the false positive detections up to 70% while maintaining a sensitivity of 95% as can be seen in figure 1.16. The International League Against Epilepsy recommends a minimum

artifact free recording time of 30 minutes [59]. Following that guideline, a routine EEG that is predicted by our model will have on average 7.5 false predictions with a sensitivity of $95_{5/8}$%. This reduction in false positives does make our model a good candidate for clinical use. Our model could function as a pre filtering tool for IED detection due to the high sensitivity. Since most artifacts are automatically rejected by the model while mostly benign variants of uncertain significance are returned as false positives, expert knowledge is required for further classification.

In conclusion, iteratively adding false positives to the training dataset does improve the performance of the IED detection algorithm significantly by reducing the false positive rate. Making this method a crucial step in the training process of (similar) classification algorithms.

Enhancing the interictal epileptiform discharge detector via GAN generated EEG segments

Introduction

As already discussed in chapter 1, the MGH clinical care dataset incorporates 88297 candidate IED's with 13262 morphologically distinguishable candidate IED's. The labels of the candidate IED's ranging between 0/8 and 8/8. The MGH clinical care dataset has also around 16 million control samples with the label 0/8, creating a highly-skewed class distribution even after the applied data augmentation. In chapter 1 we added hard examples to the dataset to increase its difficulty. Some of these hard examples are 'easy examples' for human interpreters however other hard examples such as wicket spikes, POSTS and vertex waves do tend to be falsely categorized even by (beginning) experts [48], [49], [57], [58]. Due to the morphological similarities, an increased number of samples for the closely related hard examples and candidate IED's is preferred. Collecting labeled medical data is however a complex and expensive procedure, and researchers came up with another way to enlarge a dataset called, data augmentation. If applied correctly, data augmentation may elevate model performance, providing a regularizing effect and reducing generalization error [28]–[30], [60]. When applying data augmentation, you are creating new, artificial but plausible examples, where simple augmentations such as geometric transformations and noise addition are widely adopted [31]. However, these fairly simple techniques do have a limited diversity since they heavily rely on the original data.

This lack of diversity gives incentive to a more advanced data augmentation technique called generative modeling. Generative modeling is the opposite of discriminative modeling, in which SpikeNet can be placed in. In discriminative modeling, the model tries to learn the probability of class *y* given input *x*, also known as conditional probability distribution. In generative modeling, the model tries to learn the joint probability distribution, that input data *x* and output label *y* do occur simultaneously [61]. In other words, the model learns a hidden structure of the data from its distribution and is therefore able to generate new data samples within the same distribution [34].

Various generative models are present today, including Latent Dirichlet Allocation, Gaussian Mixture Model, Restricted Boltzmann Machine, Deep Belief Network, Variational Autoencoder (VAE) and Generative Adversarial Network (GAN). Recently, the latter two do have gained the most interest due to their excellent ability of capturing key elements from a diverse range of datasets to generate realistic samples leading to sophisticated domain-specific data augmentation [30], [62].

The performance of the VAE and GAN is promising and mostly similar [35]. Comparing a VAE and GAN is subjective due to the lack of sufficient performance metrics, however some recurring performance trends are found; VAE tends to create more blurry images and are therefore lacking detail. On the contrary, a GAN usually generate sharper images and tend to be more flexible but has issues concerning training stability and sampling diversity [36], [63]. The loss of detail from a VAE in generating EEG will translate to the loss of the higher frequencies, resulting in synthesizing a slower EEG than intended, motivating the use of a GAN over the use of VAE for EEG synthesis.

Recently, an increasing number of medical studies incorporated GANs, with implementations ranging from image synthesis of the retina [64], liver lesions [30] and breast cancer tissue [37] to up sampling and synthesizing EEG [35], [65]–[67] among others. In this chapter, we investigate the applicability of various GAN's to synthesize IED's and their ability to enhance the performance of SpikeNet.

Method

During this study, we built a GAN from scratch, evaluated the performance and enhanced the GAN accordingly. The enhancement can be translated back into three major changes in the GAN algorithm which will be described later on in this chapter. Our multi-phase developmental and experimental approach which lead to our third GAN, is chosen to be described as if we are comparing the three GAN algorithms simultaneously for the sake of the readability of this chapter.

Generative adversarial network

The basic principle of a GAN

A GAN is, strictly speaking, an adversarial modelling framework for training a generative model, and was proposed by Goodfellow et al. in 2014 [62]. It is common to use deep neural networks such as convolutional neural networks in the architecture of a GAN but this is not mandatory. The architecture consists of a generator (G) and a discriminator (D). The task of the discriminator is to distinguish between real and generated data, whereas the task of the generator is creating realistic data and is therefore trying to fool the discriminator. Applying this to a real life example, the generator could be seen as a counterfeiter whereas the discriminator is the art connoisseur, where ideally the competition causes improvements in both models until the generated data is undistinguishable from the real data [68].

The generator takes a random noise vector z, sampled from a Gaussian distribution, as input and outputs fake data $\tilde{x}$ also denoted as $G(z)$. The fake data $\tilde{x}$ is passed to the discriminator together with a randomly selected real data sample, x, where they are classified as real or fake (figure 2.1).

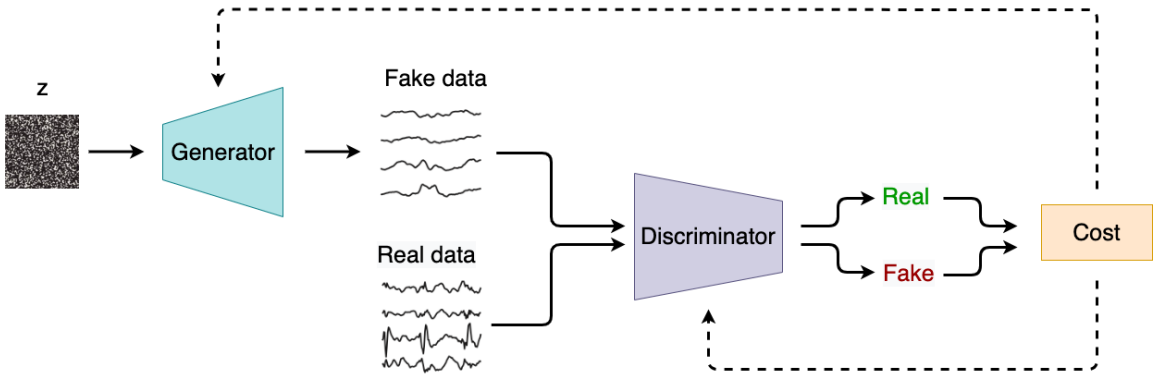

Figure 2. 1 An overview of the general architecture of a generative adversarial network. The dotted lines do represent the backpropagation to update the model parameters. Specific cost functions will be discussed later on and are therefore not incorporated in this figure.

The two-player game

Since the generator and discriminator are trained in a competitive way, the training can be seen as a two-player game with non-cooperative players. The two players, represented by the generator G, using parameters θ^G and discriminator D, using parameter θ^D, take turns on optimizing their loss function. The discriminator wants to minimize the loss function $\mathcal{L}_D(\theta^D, \theta^G)$ while changing θ^D. The generator wants to minimize $\mathcal{L}_G(\theta^D, \theta^G)$ by changing θ^G. The loss functions are defined as,

$$\mathcal{L}_D(\theta^D, \theta^G) = \mathcal{L}_D^{GAN} = -\mathbb{E}_{x\sim\mathbb{P}_r}\big[\log\big(D(x)\big)\big] - \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}\big[\log\big(1 - D(\tilde{x})\big)\big] \tag{2.1}$$

and

$$\mathcal{L}_G(\theta^D, \theta^G) = \mathcal{L}_G^{GAN} = \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}\big[\log\big(1 - D(\tilde{x})\big)\big] \tag{2.2}$$

with $\mathbb{P}_r$ and $\mathbb{P}_g$ respectively denoting the data distribution and model distribution [69]. The loss functions do partly depend on the parameters of the other player parameters leading to the description of a two player game instead of an optimalization problem [68].

Challenges during training

As addressed earlier, GAN's do have issues concerning training stability and sampling diversity.
In, for example a classification problem, the gradient of the loss is calculated, and the model parameters are optimized accordingly. Optimally, each step would lead to a lower loss which finally results in finding the global minimum of the loss landscape. In a classification problem, this loss landscape is static, however in GAN's the loss landscape changes a little every training step, making it very hard to find the global minimum in a high dimensional loss landscape, and could lead to exploding or vanishing gradients [70].

In addition to the convergence problems, GAN's can suffer from another failure mode called 'mode collapse'. During mode collapse, the generator learns to only generate a subset of all outcomes (or modes) of the data distribution $\mathbb{P}_r$. Therefore, different inputs of z lead to the same output $\tilde{x}$ [71].
Different hypothesis are presented in the literature however, to our knowledge, the true mechanisms of the mode collapse is not discovered yet.

Strategies of improvement

Wasserstein GAN with gradient penalty

The Wasserstein GAN (WGAN) uses the same adversarial modelling framework as a normal GAN however the discriminator, who normally predicts the probability of a sample being real or fake, is replaced by a critic, who predicts the realness or fakeness of a given sample by calculating the Earth-Mover (EM) distance [36]. The EM distance, or Wasserstein loss, is the minimal cost of transforming data distribution $\mathbb{P}_g$ to data distribution $\mathbb{P}_r$. Resulting in an improved stability and a meaningful loss metric [36]. Gradient penalty was proposed by Gulrajani et al. as an addition to the Wasserstein loss function, which lead to even further improvements in the stability. After implementing both the Wasserstein loss and the gradient penalty, the loss functions can be defined as,

$$\mathcal{L}_D^{WGANGP} = \mathcal{L}_D^{WGAN} + \lambda \mathbb{E}_{\tilde{x}\sim\mathbb{P}_{\tilde{x}}}[(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2] \tag{2.3}$$

With

$$\mathcal{L}_D^{WGAN} = -\mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})\,] \tag{2.4}$$

and

$$\mathcal{L}_G^{WGANGP} = -\mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] \tag{2.5}$$

Where $\mathbb{P}_{\tilde{x}}$ is defined to sample uniformly between pair of points sampled from $\mathbb{P}_r$ and $\mathbb{P}_g$ [35].

Optimalisation methods

During training, the goal is to optimize your neural network and therefore minimize the loss function. Minimizing the loss function can be achieved via various techniques. A practical and well performing technique for optimizing your network is stochastic gradient decent (SGD). SGD does yield good results with the correct parameters, however the tuning of the parameters is hard and, optimally, do need adjustment during training [72]. In response, multiple adaptive optimizer have been created including ADAM, RMSprop and Adadelta [73]. On the time of writing, ADAM is probably the most used optimizer and is recommended by Gulrajani et al. to use in the Wasserstein GAN with gradient penalty [70], [74], [75]. Recently, AdaMod, a new optimizer that builds on ADAM, was proposed and claims to be less sensitive to the chosen learning rate, to have an improved convergence and does not need a warmup [76].

Do's and Don'ts

During the years, studies have led to a better understanding of GAN's and many recommendations on how to train and build your GAN's are proposed. Radford et al. [77] proposed following architecture guideline.

> " Architecture guidelines for stable Deep Convolutional GANs
> Replace any pooling layers with strided convolutions (discriminator) and fractional-strided convolutions (generator).
> Use batchnorm in both the generator and the discriminator.
> Remove fully connected hidden layers for deeper architectures.
> Use ReLU activation in generator for all layers except for the output, which uses Tanh.
> Use LeakyReLU activation in the discriminator for all layers. "
>
> — (Radford et al.[77])

Chintala, the Co-author of the cited paper above, did give additional information about recommended implementation techniques during his presentation at NIPS [68]. Useful recommendations regarding our study are:

The use of Gaussian Latent Space instead of a uniform distribution.
Feed separate batches for real and fake to de discriminator
Use soft labels instead one hot encoding
Introduce a small percentage of incorrect labels

Implementing 3 GAN models

Dataset

The IED's that we use for this study are coming from the same routine clinical care dataset that was used as described in chapter 1. In this study we include of both medoid and member candidate IED's with the label 8/8 leading to the inclusion of 14874 candidate IED's. Choosing only one label gives us the opportunity to label the generated spikes with the same label as trained upon, if the generator is able to learn the data distribution $\mathbb{P}_r$.

Implementing GAN's

All implemented GAN's do yield the same architecture for the generator and discriminator to ensure the changes is output can be related to the optimalisations steps that are implemented.

The Generator

The generator is built to create input segments for SpikeNet, SpikeNet takes an input of 1 second EEG sampled at 128 HZ with 37 EEG channels, however the 18 bipolar montage channels can be calculated from the CAR montage. To ensure the correct relationship between the bipolar and CAR montage, it was chosen to generate the CAR montage instead of generating the CAR and bipolar montage together. Therefore, the generator is built to create 1 second epochs of EEG at 128Hz using the CAR montage.

The 19 CAR channels are generated in the following order: FP1-avg, F3-avg, C3-avg, P3-avg, F7-avg, T3-avg, T5-avg, O1-avg, FZ-avg, CZ-avg, PZ, -avg FP2-avg, F4-avg, C4, -avg P4-avg, F8-avg, T4-avg, T6-avg, O2-avg. Using the recommendations from Chintala and Radford et al. in mind the following architecture is used.

The random noise vector z with the dimension (100,1) is fed into the dense layer where it will be up-sampled to (3200,1). Reshaping this vector will give us the base of our EEG resulting in a 5 by 4 matrix with 160 filters. In the first block, which is repeated twice, the EEG will be up-sampled by the transposed convolutional layer doubling the dimensions and reducing the filters by 32. After two rounds of up-sampling, our generated EEG has a dimension of 20 by 16 with 96 filters. Since we only need 19 channels, we cut of one channel in the reshape layer resulting in an EEG segment of 19 by 16 with 96 filters.

In the second up-sampling block, the EEG length is doubled while the filters decrease with 32 per block. Resulting in an EEG sample with the dimensions of 19 by 128. The Tanh function does scale the EEG between -1 and 1, to compensate for this, the EEG is multiplied by 500 creating an EEG in the range of 500 and -500 µV.

The Discriminator

To maximize the similarities between the generator and the discriminator, which may lead to a more stable training, a mirrored architecture of the generator is used. Here the first convolutional block will down-sample the length of the EEG by 2 and increase the filters by 32. The second block will reduce the height and width of the EEG by 2 and will increase the filters by 32. Passing all convolutional block will lead to a matrix of 5 by 4 by 160, which is the same size as the starting point for the generator. Where finally the prediction real and fake is made by the Leaky ReLU.

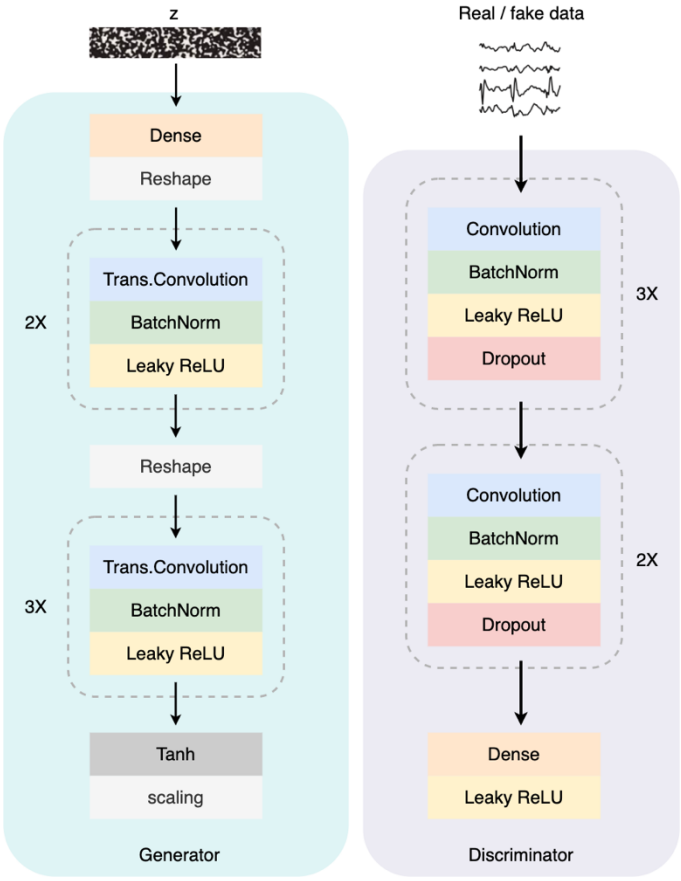

Figure 2. 2The architecture of both the generator in blue and the discriminator in purple. The architectures are created such that they are very similar regarding the data size and number of filters.

The shown versions

During this study, many experiments are conducted but not all will be shown. The experiments can be categorized in within three groups accounting for the major changes. The models that we will show are:

The GAN + ADAM optimizer (GAN-ADAM)
The Wasserstein GAN with gradient penalty + ADAM optimizer (WGANGP-ADAM)
The Wasserstein GAN with gradient penalty + Adamod optimizer (WGANGP-Adamod)

Evaluating GAN's

Evaluating generated data is challenging, since multiple answers can be correct and only the realness of the data needs to be evaluated. Human interpretation ceases to be a main evaluation metric in the beginning of the GAN's. Over the years evaluation metrics are proposed such as the widely adopted Inception Score (IS), Frechet Inception Distance (FID) and Euclidean Distance (ED). The first two do rely on a pre-trained image classification model, requiring a square input and judging realness based on image features. It is not hard to imagine that these metrics will not produce useful or even reliable scores when applied on EEG.

Calculating the ED is not able to tell us how real or unreal our generated samples are; however, it can tell us if the model re-produces samples from the input domain $\mathbb{P}_r$ and is therefore used in our evaluation. In addition to the ED we evaluate if our generator does produce IED's, which is our main goal. We accomplish this by generating 10.000 IED's at the end of each training epoch and feed them into SpikeNet. We monitor the total number of detected IED's by SpikeNet as well as the average outcome of the 10.000 IED. An increase in those scores, which are heavily related, will give us insight in the performance of the generator. Both scores will not give us any insight if mode collapse is present, therefore manual inspection is also applied.

Enhancing SpikeNet with the generated spikes

Enhancing SpikeNet by adding the generated IED's to the dataset with the label 8/8 might look like the obvious approach. If our best performing GAN produces IED's, that truly belong in the data distribution of the label 8/8, 100% of the time, the latter approach will be useful. However, it is most likely that our GAN will not be able to produce 8/8 IED's all the time, making automatic labeling impossible without incorrectly labeling some of the generated IED's. To overcome the problems of labeling the generated IED's, we make one assumption.

Looking at the results of chapter 1, we assume that the only difference between SpikeNet in training iteration 1 and 15 is the false positive rate. Using that assumption, we generate IED's, predict the IED's with both the SpikeNets from iteration 1 and iteration 15. Dividing the prediction SpikeNet_{15} from prediction SpikeNet_1, gives you information about the likelihood of the given sample being a false positive or not. Values close to 0 are likely to be real where values close to 1 are likely to be a false positive.

Adding the generated IED's as 'false positives' to the dataset with the label 0/0 might lead to better performance of SpikeNet. The distribution of the outcome difference, between SpikeNet_1 and SpikeNet_{15}, as shown in figure 2.3, is a bell curve with one long tail. Based on this distribution, which is calculated on 100.000 generated IED's, it is chosen to include generated IED's with a where the outcome of SpikeNet_1 minus the outcome of SpikeNet_{15} is greater than 0.4.

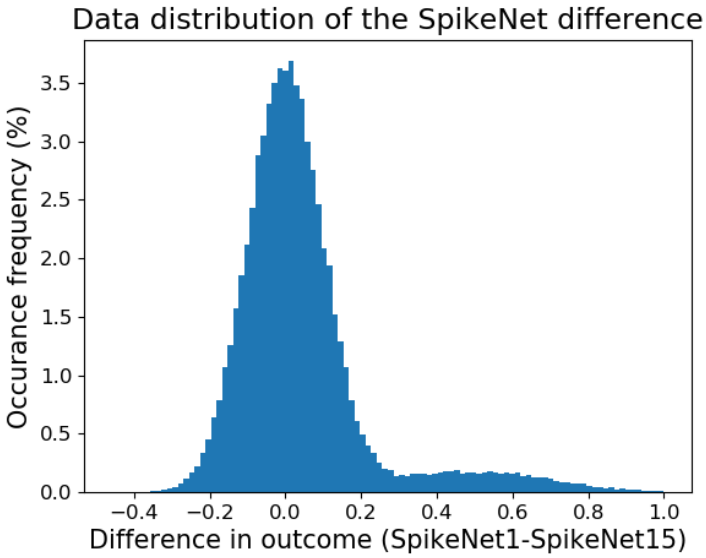

Figure 2. 3 The data distribution of the difference in outcome between SpikeNet1 and SpikeNet15. 100.000 samples are generated and are predicted by both SpikeNets. The right tail is longer than the left tail the the bell curve indicating the presence of generated false positives.

Training procedure

All models are initially trained for 1000 epochs, early termination of the training progress will be applied if the model if no indication of improvement is present, while the model fails to converge, suffers from significant mode collapse or generates IED's with morphologies far from actual IED's.

Results

Convergence

During the training of the three models, which are all trained multiple times, the convergence is the first and easiest thing to evaluate. After each training step, the generator and discriminator loss are calculated and visualized. When the losses do converge to zero, it indicates that GAN is finding an equilibrium between the generator and discriminator, resulting in a balanced training.

GAN + ADAM optimizer

The GAN-ADAM is our most simple implementation of the 3 models, and as shown below in figure 2.4, and is not able to have a stable training process.
The orange line shows the generator loss and the blue line the discriminator loss. As seen no convergence is present.

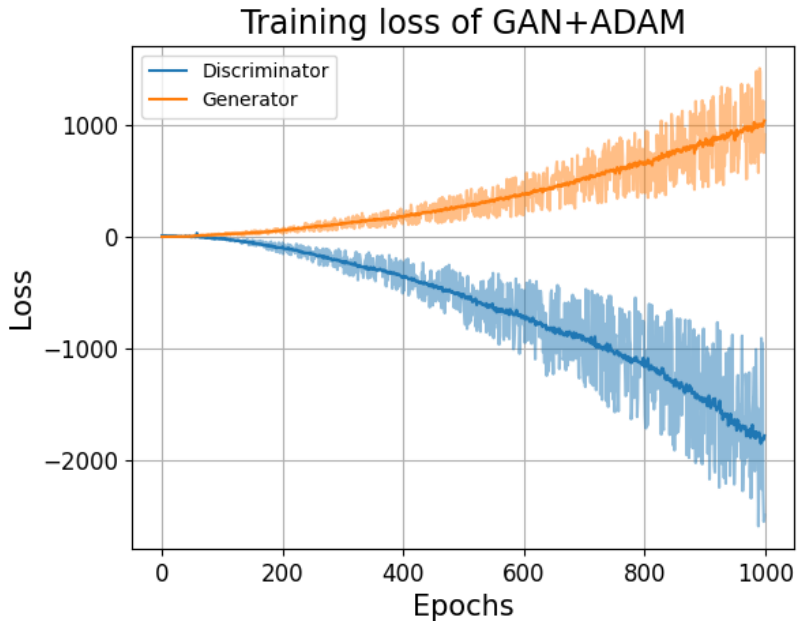

Figure 2. 4 The loss plotted during a typical training of the GAN with the generator loss in orange, discriminator loss in blue. As seen, both the losses do diverge from resulting in an unstable GAN.

The Wasserstein GAN with gradient penalty + ADAM optimizer

The WGANGP-ADAM is known to enhance the stability of the GAN and our implementation did achieve a higher stability. Even though some runs did fail to converge, the majority did, as shown below. In the majority of the runs an oscillating motion is seen in the generator loss (orange curve), which indicates that mode collapse is present. The oscillating motion is created when the generator switches from generating one mode to another. Different modes do yield different losses however the generator is not able to produce multiple modes simultaneously. As seen, for the first ~50 epochs the discriminator loss (in blue) seems to be in a free fall. Most likely finding a way to discriminate between real and fake samples. As seen, the generator loss stays practically zero. When the discriminator finds a way to discriminate between real and fake, the generator will start learning useful features.

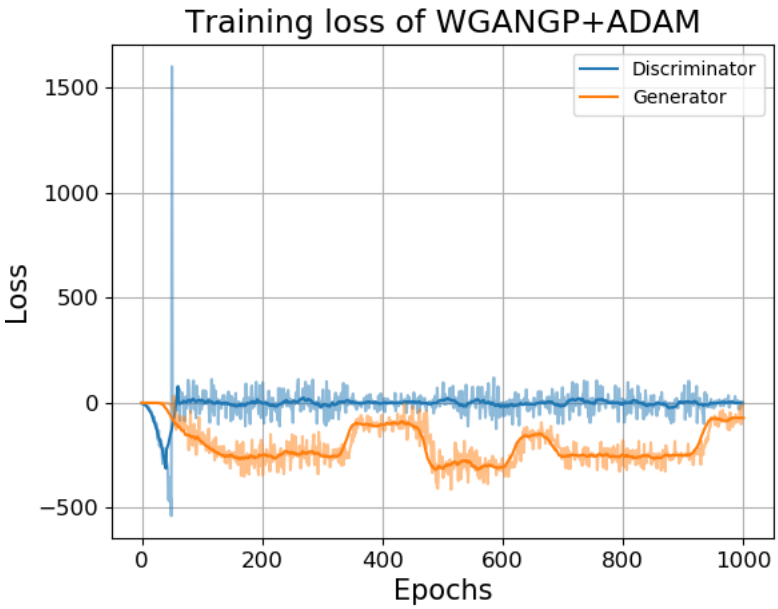

Figure 2. 5 A typical training of the WGANGP+ADAM. In the first epochs, the discriminator (blue) is not able to discriminate between real and fake, hence the downward slope. When de discriminator finds a way to discriminates the generator start learning useful features. The plateaus seen in the generator loss (orange) are a hallmark of mode collapse. Each plateau is created when the generator only outputs a specific a subset of the classes, where different plateaus are created by different subsets.

The Wasserstein GAN with gradient penalty + Adamod optimizer

The Adamod optimizer does very strictly what it promised to do in the WGANGP-Adamod configuration. It increases the stability and the learning ability in the beginning, however over time, it is not able to hold the equilibrium. As can be seen in figure 2.6, the discriminator loss slowly deviates while the standard deviation of the discriminator loss increases. Due to the stable results it was chosen to additionally train the WGANGP-Adamod for more epochs to see if it leads to better results.

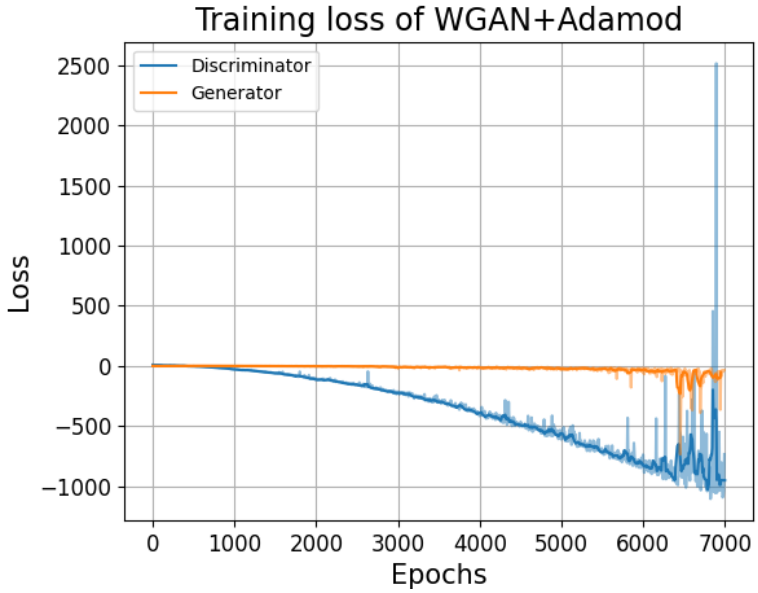

Figure 2. 6The loss plotted during a typical training of the WGANGP-ADAM with the generator loss in orange and the discriminator loss in blue. The gradient penalty is given in light blue

ED score

The Euclidean distance is calculated between all input data and 10.000 generated IED's per generator. The more the generated IED's mimics an input IED, the lower the ED score. The ED score is the average, point wise, Euclidean distance between the generated IED's and the input IED's. To give an indication what the mean ED should be, the ED score is calculated on itself using the input IEDs.

Table 2. 1 The ED score of the best performing generator of each catagory. The ED score is also calculated on itself to create a reference value

	GAN+ADAM	WGANGP+ADAM	WGANGP+ADAMOD	Input IEDs
Minimal ED	281.1	9.7	86.9	0.1
Mean ED	306.0 ± 8.4	52.8 ± 17.4	117.0 ± 9.9	39.6 ± 6.3

Visual evaluation

Before each training, 9 latent vectors are saved for evaluation purposes. At the end of each epoch, the 9 latent vectors are fed into the generator, creating 9 IED segment. In this way, the development of the IED's could be tracked. After the training was finished, more IED segments are evaluated to check for mode collapse. Each segment is plotted as a one second, 19 channel CAR montage as described earlier.

GAN

The instability issues of the GAN-ADAM do lead to improper training of the generator resulting in a noisy output ranging between -500 and 500 uV as seen below in figure 2.7.

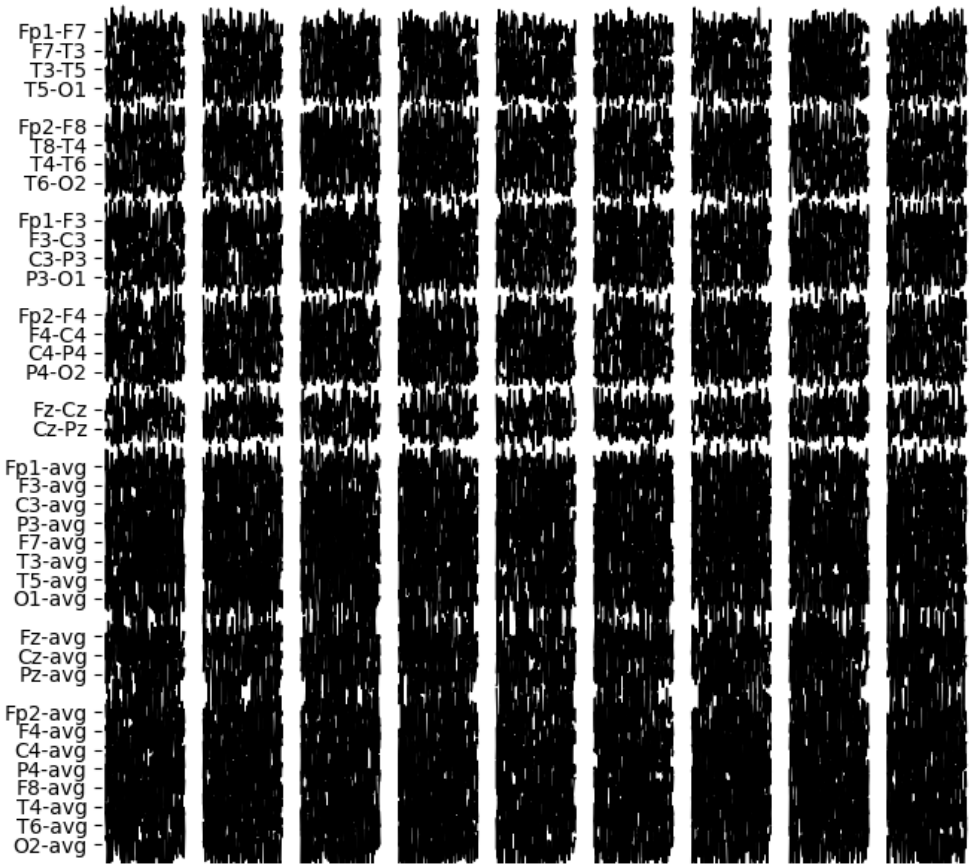

Figure 2. 7 Nine one-second generated IED segments created by the GAN-ADAM are shown. The failure to converge lead to the output of a noisy signal.

The Wasserstein GAN with gradient penalty + ADAM optimizer

The majority of the WGANGP-ADAM models suffered from mode collapse during training, however some models did avoid mode collapse. Both outcomes, are presented to give an insight in the generated samples.

In figure 2.8 the outcome of a WGANGP-ADAM model with mode collapse is shown. As seen, a high similarity between the first and last three segments, and between segment 4 and 5 is present. Looking at the wave forms, all segments suffer from an upward slope at the end also, the morphology of the generated segments does not come close to an IED except for segment 6.

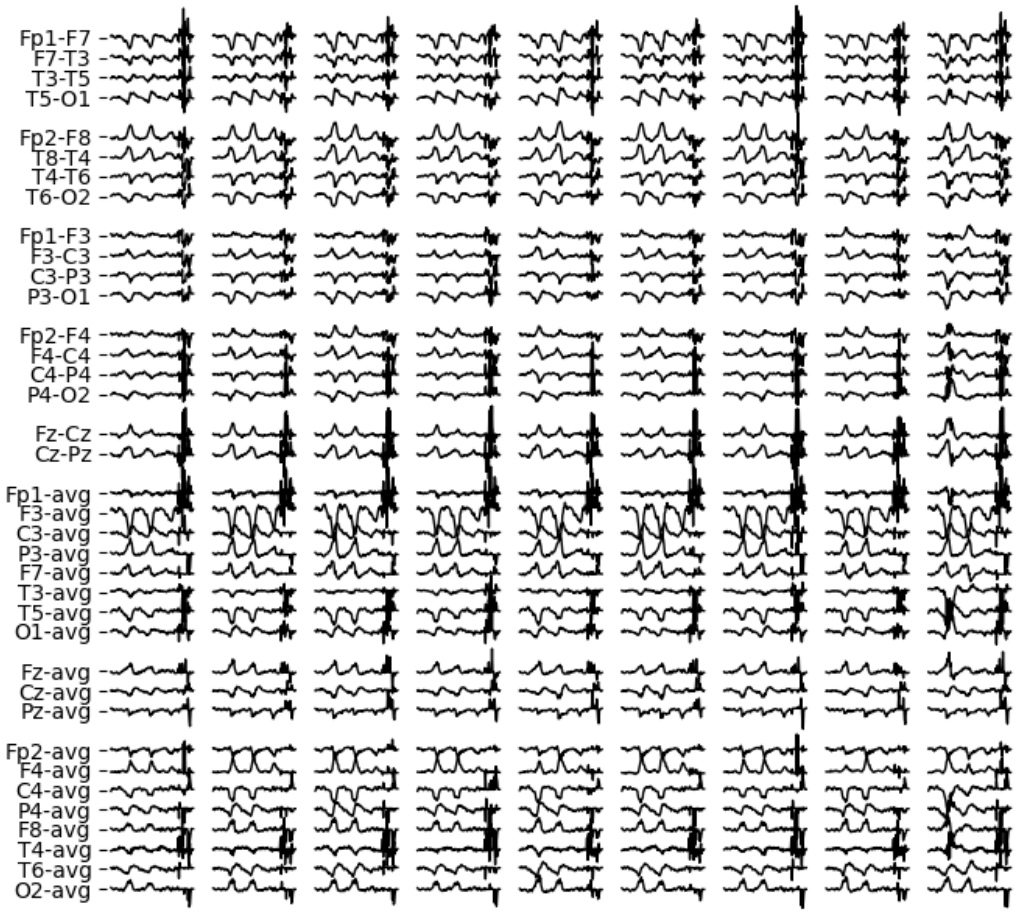

Figure 2. 8 Nine one-second generated IED segments created by the WGANGP-ADAM are shown while mode collapse is present.

If the WGANGP-ADAM did not suffer mode collapse, the generated IED's do not only yield more variance but also a show a higher voltage EEG and a less present upward slop in the end. It can be seen that some spike like behavior is learned however the morphologies are in most cases vastly different from real IED's.

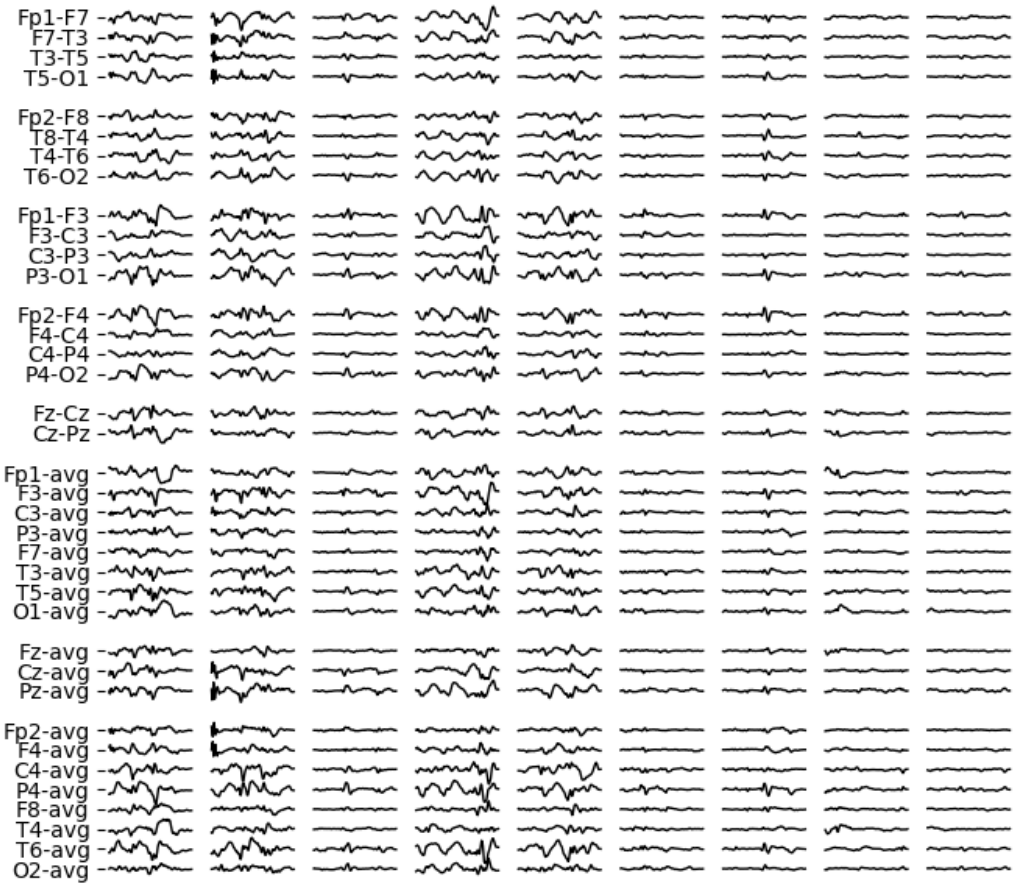

Figure 2. 9 Nine one-second generated IED segments created by the WGANGP-ADAM are shown without a noticeable level of mode collapse.

The Wasserstein GAN with gradient penalty + Adamod optimizer

The increased performance lead to longer training sessions which on its turn lead inevitably to mode collapse. However, before mode collapse occurred generators are present which do have an improved outcome regarding the morphology of the IED as shown below in figure 2.10. As seen, in the majority of the samples, higher voltages are present in the lower half of the channels. Even though this similarity, many differences are found between the generated IED's. Overall it could be stated that the generated IED's of the WGANGP-Adamod do yield a more pronounced spiky pattern leading to morphologies with a higher tendency towards IED's.

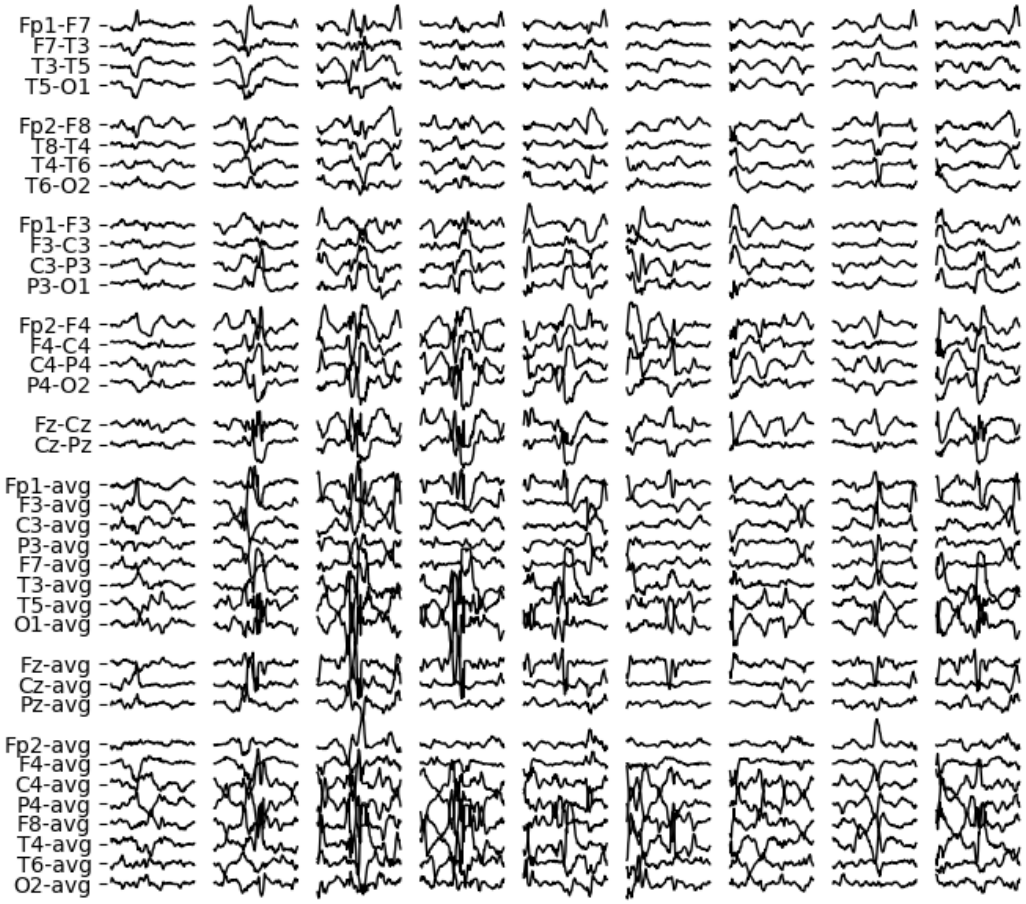

Figure 2. 10 Nine one-second generated IED segments created by the WGANGP-Adamod are shown. No mode collapse is present in tis figure..

SpikeNet evaluation

After each epoch the generator was evaluated by $SpikeNet_{15}$. Only the WGANGP-Adamod succeeded in generating IED's that where classified as IED more than 60% of the time for multiple consecutive epochs, without suffering from mode collapse. WGANGP+ADAM does reach the high percentages but only happens in severe mode collapse.

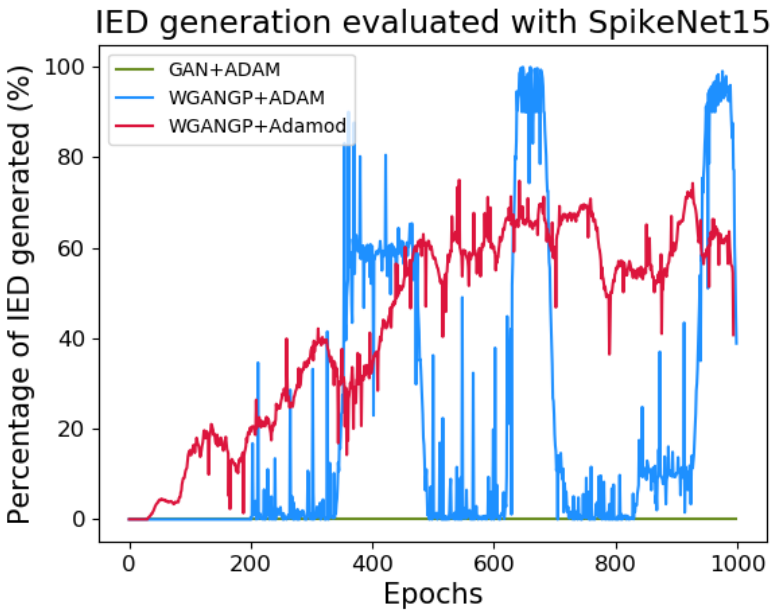

Figure 2. 11 The percentage of generated IED's evaluated with SpikeNet 15. At the end of each epoch, 10.000 IED's are generated and evaluated with SpikeNet15. Based on the outcome of SpikeNet15, the GAN+ADAM is not able to generate IED's, WGANGP+ADAM Is able to generate IED's but suffers severe mode collapse and WGANGP+Adamod learns to produce IED's up to 78% of the time.

SpikeNet enhancement

For comparison, all previous training iterations are shown. Training iteration 16, the last iteration, is the iteration where generated IED's are appended to the dataset as false positives.

The data shown below is created by adding generated IED's classified as false positives with minimal difference between $SpikeNet_{15}$ and $SpikeNet_1$ of 0.4 Improvements over the $AUCROC_{ajust}$ are present in all calculations except the one using the labels 7/8 or higher. However, the $AUCPRC_{ajust}$ slightly decreases in all calculations.

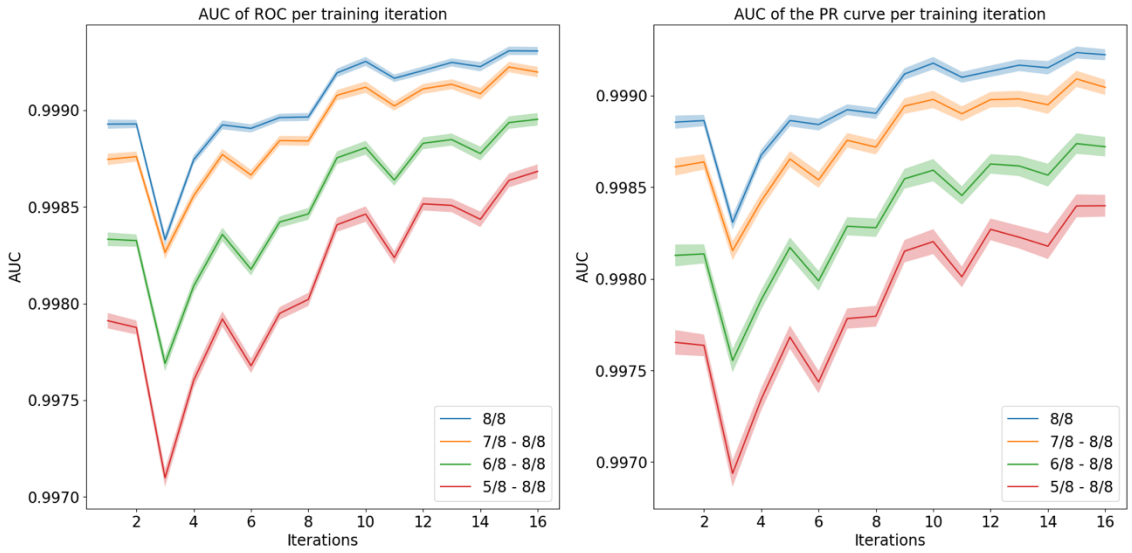

Figure 2. 12 The AUCROC and the AUCPRC plotted against the training iteration, in the 16th iteration the generated IED's are added to the dataset

Looking at the false positives per hour, plotted for the sensitivity levels of $99_{5/8}$%, $98_{5/8}$% ,$95_{5/8}$%, $99_{8/8}$%, $98_{8/8}$% and $95_{8/8}$% it can be seen that major improvements are made at $99_{5/8}$% and $98_{5/8}$% however all other parameters do have similar to worse performance.

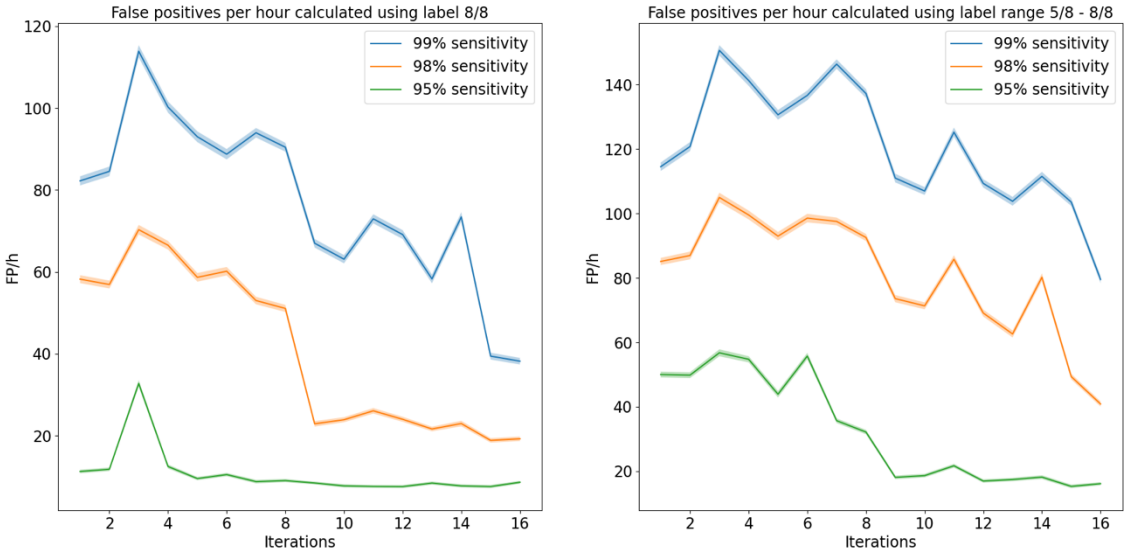

Figure 2. 13 The false positive predictions plotted against the training iterations

When the ROC of $SpikeNet_{15}$ (shown in the dotted lines) is compared with the ROC of $SpikeNet_{16}$. It can be seen that ROC of $SpikeNet_{16}$ has slightly moved upwards and to the right, resulting in a lower performance below ~19 FP/h and an increase performance above the ~19 FP/h.

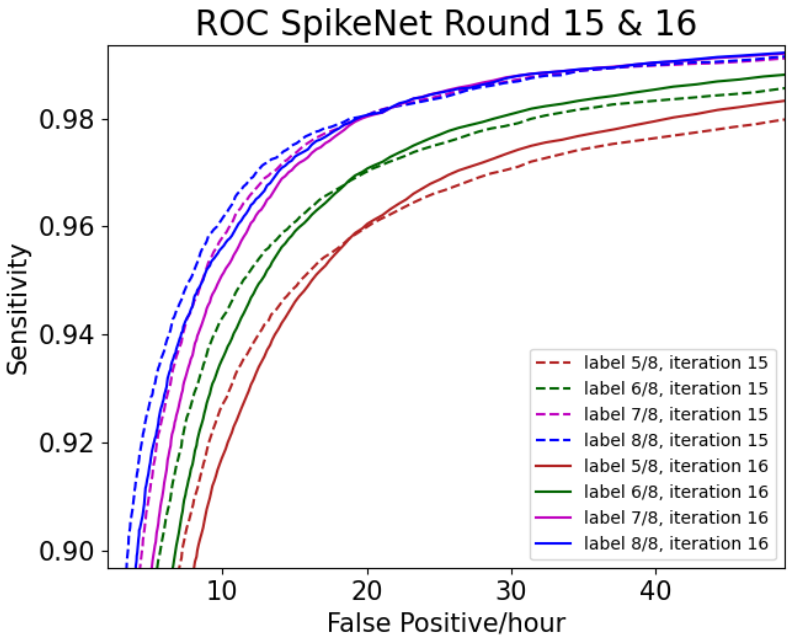

Figure 2. 14 The ROC Curve of $SpikeNet_{15}$ shown by the dotted lines and $SpikeNet_{16}$ in the continuous line. $SpikeNet_{16}$ outperforms $SpikeNet_{15}$ at FP/h rates of 19 and above. Below the FP/h rate of 19, the $SpikeNet_{15}$ is the superior one.

Discussion

During this study we showed that stacking multiple stability improvement methods does cause an increased stability and there a performanceof the GAN. We were able to go from generating noise, using the GAN-ADAM, to generating IED's with which are predicted to be an IED by SpikeNet$_{15}$ more than 60% of the time without suffering from mode collapse using the WGANGP-Adamod. It remains unclear how to use the generated IED's for performance enhancement of SpikeNet while focusing on the reduction of the FP/h since adding the generated IED's increased the 95%$_{5/8}$ from 15 to 18.3 FP/h.

The potentials of generating EEG using GANs is a fairly unexplored field, where a hand full of papers exploring this opportunity are present at the time of writing. Existing studies showed the increased performance of classification tasks when generated EEG, without the presence of pathologies, was added to the dataset [35], [65], [66]. Where Luo et al. addressed the up-sampling of EEG from 125Hz to 250Hz [67]. To our knowledge, we are the first to generate and also evaluate the generation of IED's using multiple GAN implementation.

Our work is mainly explorative and does include empirical experiments and trial and error analysis. Unfortunately, we were not able to explore all promising features, techniques and/or methods. Unaddressed issues will be discussed in the recommendations. Besides the limit of the implemented methods, our study does have drawbacks including the lack of a sufficient evaluation metric. Using the prediction from SpikeNet $_{15}$ and the ED as a metric has two downsides. Firstly, it is known from chapter 1 that SpikeNet $_{15}$ does suffer from false positives. Creating IED like segments can lead to false detections of SpikeNet $_{15}$. False detections of SpikeNet $_{15}$ do mostly include artifacts that mimic IED's and benign variants of uncertain significance like vertex waves, POSTS and wickets spikes. We can therefor say that an increased prediction of SpikeNet $_{15}$ is associated with an increased IED like behavior. Secondly, the metrics based on SpikeNet $_{15}$, are not a sufficient to detect mild or moderate mode collapse. Severe mode collapse, results in a small range of predictions. When the generator, while suffering from mode collapse, transforms from generating non-IED like segments to generating IED like segments, the percentage of predicted IED's, as classified by SpikeNet $_{15}$, will rapidly increase and will show oscillating behavior with can be linked to the generator loss when one mode changes to another. Introducing the standard deviation of SpikeNet $_{15}$ prediction as a measurement of mode collapse may look like the obvious solution. However, this will only work if, and only if, one mode is generated, mode collapse with two vastly different modes can still result in a high standard deviation and is therefore chosen not to be implemented.

The strength of our study lies within the explorative and validating nature. We explored the field of generating IEDs by using widely adapted deep neural network techniques which proved their performance in other fields [36], [62], [70], [76]. By systematically changing parameters we are able to show the individual changes that our different implementations provided. Henceforth, creating a promising starting point for further research. Even though optimalisation is preferred, our approach of adding generated IED's did overcome the problem of labeling the patients and enabled us to add the generated IEDs to the dataset based on our domain knowledge.

During the evaluation of SpikeNet$_{16}$ it was found that the ROC moved upwards and to the right, compared to the ROC of SpikeNet$_{15}$, as seen in figure 2.14. The ROC values from 19FP/h and up do indicate that the model performance is increased. In other words, reducing the false positive predictions while maintaining the sensitivity. Meaning that the difference in model prediction between false positives and IEDs has increased. This is especially true in the area of low predictions (SpikeNet outcome of <0.28), leading to a higher distinguishability of the IED's, which is good. However, using higher thresholds than 0.28 (current is 0.43), the model prediction slightly drops leading to a decreased performance. In other words, the added IEDs caused a reduced distinguishability between false positives and spikes at higher thresholds. This performance increase can be explained by the fact that

a part of the generated IED's indeed where false positives, leading to the improved performance at low thresholds. However, the reduction in performance can be caused by two mechanisms. Firstly, it could be that some of the generated IED's might not be an actual false positive, creating high gradients in the backpropagation which lead to a lower prediction of the IED's. Secondly, all the generated IED's where indeed false positives, however some have a highly similar morphology compared to the IED. For example, the model generated benign variants of uncertain significance. Due to the high morphological similarities, SpikeNet was not able to learn the fine-grained differences resulting in a performance drop.

In conclusion, we implemented three versions of our IED generating GAN, where the WGANGP-Adamod, who was designed to be the most stable, outperformed all other models. Even though a suitable way to include the generated IEDs in the training process of SpikeNet is not found, the promising results regarding IED generating without suffering from mode collapse, do pave the way for future work.

Validating and enhancing the (temporal) Brain Symmetry Index

Introduction

The EEG is a complex interplay of frequencies that is created by the underlaying cortex.
When cerebral disfunction occurs, the frequencies created of the affected cortex slow down, also known as background slowing [27], [57]. The EEG can provide evidence of underlying cerebral disfunction by demonstrating this background slowing. The slowing of the background activity can be presented focally as well as generalized, depending on its underlying cause. However, generalized background slowing cannot immediately be linked to cerebral disfunction since generalized background slowing in the theta (4-8Hz) and delta (0.5-4Hz) range of the EEG can represent developmental slowing, during drowsiness and sleep and it can be induced by sedative medication. However, intermittent generalized slowing or persistent focal slowing of a specific region, as well as unreactive focal or generalized slowing should be considered pathologic [27]. Pathologic causes of generalized slowing include encephalopathies[78], neurodegenerative disorders[79], as well generalized anoxia caused by a cardiac arrest[80]. Focal slowing can be caused by numerous underlying mechanisms like, ischemic stroke[81], brain hemorrhage[82] and tumors[83].

As earlier described, visual scoring of the EEG is still the golden standard, even though multiple automated EEG analysis algorithms are present. Van Putten et al.[84] published in 2006 a method to detect the occurrence of generalized slowing and asymmetry in the EEG, called temporal brain symmetry index (*tBSI*) and brain symmetry index (BSI) respectively. In 2007 Van Putten published a refined version of his previously published slowing asymmetry detection called the revised temporal brain symmetry index (*r-tBSI*) and revised brain symmetry index (*r-BSI*) respectively[85], [86].

Both the *tBSI* and the *BSI* where initially designed to assist detecting feature changes in the EEG, such as detecting (partial) slowing of the EEG during carotid endarterectomies. Both yield promising results. However, there is one major drawback in the approach of the *tBSI*. A design choice of the *tBSI* is the dependency of a reference epoch coming from the same EEG. Already present continuous slowing cannot be detected since the slowing will be present in the reference epoch, making the *tBSI* in combination with the proposed method unsuitable for detecting continuous slowing. In contrast to the *tBSI*, the *BSI* does not depend on a reference epoch and detection is not affected.

The dependence on a reference does not necessarily have to be changed before generalized slowing can be detected. The reference power calculated using the patient's own EEG could be replaced by a general reference. A general reference could overcome the burden dependency of the same EEG. This reference matrix must include the power spectral density (PSD) of the EEG per channel per age. The general reference must at least include the PSD per channel, however the PSD per channel per age is highly recommended since the PSD of the EEG has been proven to changes during aging [22], [87], [88].

In this study, we calculate a general reference matrix which includes the PSD per channel per age.
We implement both the *r-tBSI* and *r-tBSI* using the general reference matrix. Subsequently, we validate both the original and revised *tBSI and BSI.*

Method

Data preparation

For the creation of the general references, all normal EEG's (n=4656) are included. All EEG's are resampled to 128Hz and are high pass and notch filtered using 0.5Hz and 60Hz accordingly.
All EEG's are constructed to be in the anterior-posterior montage, also known as double banana (DB) montage. For each bipolar derivation, the PSD is calculated using a 5 second non overlapping window and is stored accordingly. Creating a Nx321x18 matrix with N the number of 5 second epochs.

Creating the general reference matrix via averaging per age

The PSD, with a frequency resolution of 0.2 Hz, is calculated for all 5 second epoch, to ensure that no asymmetry is present in the general reference matrix, the PSD's of the following channel combinations are averaged: FP1-F7 & FP2-F8, F7-T3 & F8-T4, T3-T5 & T4-T6, T5-O1 & T6-O2, FP1-F3 & FP2-F4, F3-C3 & F4-C4, C3-P3 & C4-P4 and P3-O1 & P4-O2. After averaging per channel, the PSD's are averaged per age. The reference at age *y* is derived by calculating and averaging the PSD for all patients with the age of *y*-1*, y* and *y*+1. Using a bandwidth of 3 years will smooth the PSD while retaining ability to capture the changing PSD. The main function of this PSD matrix is to enable visual evaluation of the deep learning created reference matrix.

Creating the general reference matrix via a deep learning model

No complex smoothing techniques are used in the general reference matrix which is described above. To create a smoother reference matrix, a deep learning model used. To ensure the symmetry of the reference matrix, the same channel combinations are averaged as mentioned above, leaving 8 unique EEG channels. For each unique EEG channel, a deep learning model was trained based on Sun et al.'s Deep Neural network. The model is trained to predict the spectrogram $S_{i,f,c,j}$ on a log scale (Decibel), where f is the frequency; c the channel; and j the epoch within patient i. The model takes $[a_i, a_i^2, a_i^3, a_i^4, a_i^5]$, with a_i the age of subject i, as input and has one hidden layer with 100 nodes.

Patient selection

Out of the available 5698 abnormal EEG's, 100 EEG's with slowing between the age of 18 and 80 were randomly chosen. It was made sure, by reading the EEG reports, that all 100 patients were diagnosed with generalized 'theta' or 'delta/theta' slowing without the occurrence of conditions that significantly change the PSD of the EEG such as: (partial) seizures, extensive muscle artifacts, breach rhythm or medication induced slowing. If one of the conditions was present, the corresponding EEG was replaced by another randomly chosen EEG until all 100 EEG's fulfilled the criteria.

For the asymmetry group, 100 patients were selected in the same manner as for the slowing group. Only now, all patients with reported 'background asymmetry' within the range of 18 and 80 years old where included.

The control groups contain 100 EEG's each with patients between the age 18 and 80. The patients were randomly selected and where checked by hand to see if slowing or asymmetry was present respectively. If present, the patient was rejected for the control group and a new randomly chosen patient was inspected until 100 patients per control group was reached.

Table 3. 1 Demographic features of the 4 patient groups.

	Slowing	Asymmetry	Control slowing	Control asymmetry
Age ± std (years)	52.1 ± 17.8	52.7 ± 16.7	46.3 ± 17.2	51.4 ± 17.9
Length ± std (min)	58.5 ± 16.1	57.9 ± 31.9	60.0 ± 22.4	56.3 ± 25.7

Implementing the BSI & r-BSI

The *BSI* and the revised *BSI* (*r-BSI*) are defined by Van Putten as a measure for interhemispheric spectral symmetry. The BSI is defined as,

$$BSI = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{M}\sum_{j=1}^{M}\frac{R_{i,j} - L_{i,j}}{R_{i,j} + L_{i,j}}\right| \tag{3.1}$$

with $R_{i,j}/L_{i,j}$ the power calculated at the right/left hemispheric bipolar derivation *j* while using the frequency range *i*.

While the *r-BSI* can be written as,

$$\text{r-BSI(t)} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{R_i^*(t) - L_i^*(t)}{R_i^*(t) + L_i^*(t)}\right| \tag{3.2}$$

with

$$R_i^*(t) = \frac{1}{M}\sum_{j=1}^{M} r_i^2(j,t) \tag{3.3}$$

and

$$L_i^*(t) = \frac{1}{M}\sum_{j=1}^{M} l_i^2(j,t) \tag{3.4}$$

where $r_n(j,t)$ and $l_n(j,t)$ are the power at the frequency *i* for channel *j* evaluated at time t for respectively the right and left hemisphere.

Implementing the tBSI & r-tBSI

The *tBSI* and the revised *tBSI* (r-*tBSI*) are created for temporal changes in the EEG. The original *tBSI* is defined as,

$$tBSI = \frac{2 \times tBSI' + BSI}{2} \tag{3.5}$$

with

$$tBSI' = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{K}\sum_{j=1}^{K}\frac{S_{i,j} - S_{ref\ i,j,a}}{S_{i,j} - S_{ref\ i,j,a}}\right| \tag{3.6}$$

where $S_{i,j}$ is denoting the power of bipolar derivation *j* calculated using the frequency range *i*. $S_{ref\ i,j,a}$ is the reference power for bipolar derivation *j* using frequency range *i* at age *a.*

While the r-*tBSI* is defined as,

$$\text{r-}tBSI = \sqrt{|(\Delta R(t) - \gamma)\cdot(\Delta L(t) - \gamma)|} \tag{3.7}$$

with

$$\Delta R(t) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{R_i^*(t) - R_i^*(t_{ref})}{R_i^*(t) + R_i^*(t_{ref})}\right| \tag{3.8}$$

and

$$\Delta L(t) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{L_i^*(t) - L_i^*(t_{ref})}{L_i^*(t) + L_i^*(t_{ref})}\right| \tag{3.9}$$

where t_{ref} is denoting the power at the frequency *j* for channel location i evaluated at the reference matrix and γ is an offset correction factor.

Implementation phase

During the implementation phase, the *BSI, r-BSI, tBSI* and *r-tBSI* are implemented as described by Van Putten et al. except that we use a total frequency range between 0.4 and 20Hz instead of 1 to 25Hz. Lowering the upper frequency range will result in lower sensitivity for EMG artifacts while retaining the sensitivity [86]. Subsequently, the reference matrices ware created and the *BSI, r-BSI, tBSI* and *r-tBSI* where predicted. During the prediction, the PSD is calculated using a 5 second non overlapping window. After implementation the performance is evaluated via the AUCROC. The original and revised algorithms are compared, and the best performing algorithm is chosen for the experimental phase.

Results implementation phase

Creating the reference matrix

Before implementation of the algorithms, the reference matrices are created. In figure 3.1 the power spectral density is plotted for the F3-C3 and F4-C4 bipolar derivation. The upper panel shows the reference matrix created by the deep learning model where the lower panel shows mathematical one using a 3-year bin. As seen the reference matrix created by the deep learning model is smoother compared to the mathematical reference matrix but follows the overall trend which is present in the mathematical model.

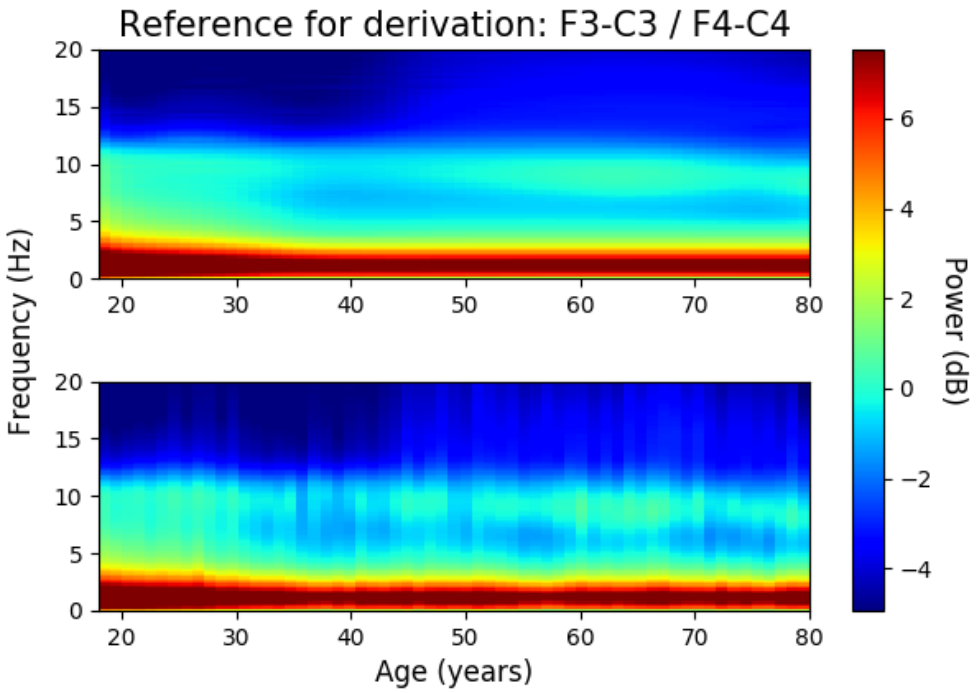

Figure 3. 1 Reference matrices. The upper panel shows the reference matrix created by the DNN. The lower panel shows the calculated reference matrix. As seen, the upper panel is the smooth version of the lower panel.

Implementing the algorithms

After implementing *BSI, r-BSI, tBSI* and *r-tBSI* the means and standard deviation per group as well as the AUCROC are calculated and shown in table 3.2.

Table 3. 2 Performance of the BSI, r-BSI, tBSI and r-tBSI.

		Mean prediction ± std (patients)	Mean prediction ± std (controls)	AUCROC	Sensitivity at 99% Specificity	Sensitivity at 95% Specificity
Asymmetry detection	***BSI***	0.346 ± 0.118	0.205 ± 0.040	**0.92**	**68%**	**75%**
	r-BSI	0.518 ± 0.160	0.268 ± 0.077	0.90	59%	74%
Slowing detection	***tBSI***	0.526 ± 0.094	0.442 ± 0.073	**0.84**	46%	**70%**
	r-tBSI	0.744 ± 0.084	0.701 ± 0.076	0.82	**56%**	65%

Experimental phase 1

During the experimental phase, changes are made compared to the original implementation.
The frequency range *i* over which the power is calculated is consists of 20 bins in the original implementation. This is reduced to 4 bins corresponding the delta (0.4 – 4 Hz), theta (4 – 8 Hz), alpha (8 – 13 Hz) and beta (13 – 20 Hz) power of the EEG and lastly it is reduced to 2 bins corresponding to the delta-theta (0.4 – 8 Hz) and alpha-beta (8 – 20 Hz) range.
The best asymmetry and slowing detection algorithm will be selected based upon the AUCROC.

Table 3. 3 Overview of experimental calculations during Experimental phase 1 with the corresponding performance.

	Number of freq. bins	Description of frequency bins	Mean prediction ± std (patients)	Mean prediction ± std (controls)	AUCROC	Sensitivity at 99% specificity	Sensitivity at 95% specificity
BSI	2	0.4Hz-8Hz (Delta-Theta) 8Hz-20Hz (Alpha-Beta)	0.524 ± 0.171	0.236 ± 0.107	0.94	44%	63%
	4	0.4Hz-4Hz (Delta), 4.2Hz-8Hz (Theta), 8.2Hz-13Hz (Alpha), 13.2Hz-20Hz (Beta)	0.346 ± 0.118	0.205 ± 0.040	**0.95**	**72%**	**77%**
	20	0.4Hz-1Hz, 1.2Hz-2Hz, 2.2Hz-3Hz, … 19.2Hz-20Hz	0.343± 0.119	0.204± 0.039	0.92	68%	75%
tBSI	2	0.4Hz-8Hz (Delta-Theta) 8Hz-20Hz (Alpha-Beta)	0.600 ± 0.139	0.691 ± 0124	0.69	12%	22%
	4	0.4Hz-4Hz (Delta), 4.2Hz-8Hz (Theta), 8.2Hz-13Hz (Alpha), 13.2Hz-20Hz (Beta)	0.526 ± 0.094	0.442 ± 0.073	**0.88**	**55%**	**76%**
	20	0.4Hz-1Hz, 1.2Hz-2Hz, 2.2Hz-3Hz, … 19.2Hz-20Hz	0.519 ± 0.094	0.453 ± 0.074	0.84	46%	70%

Experimental phase 2

When looked closely at the *BSI,* it can be seen that the outcome is based on the relative difference between the channels. This can give misleading values while the power approaches zero, for example, in the beta range during extreme slowing, it will also be misleading when the power of the total signal approaches zero during burst suppression or in an isoelectric EEG.

These circumstances have one thing in common, the calculated power at $R_{i,j}$ $(L_{i,j})$ is far below healthy value. The smaller the calculated power, the more influence an unsignificant absolute differences has. To overcome this, a correction factor is created which starts at 1 and will approach zero when de calculated power deviates from the healthy value.

The correction factor needs to take the power of both hemispheres into account. To do so, the percentual slowing between $R_{i,j}$ and $L_{i,j}$ is calculated which can be derived as follows. The absolute *BSI* at frequency *i* and derivation *j* is defined as

$$abs\left(BSI_{i,j}\right) = \left|\frac{R_{i,j} - L_{i,j}}{R_{i,j} + L_{i,j}}\right| \tag{3.10}$$

Rewriting formula 3.10 gives us the relative percentage of slowing between the derivations, defined as

$$P_{min} = \frac{\min\left(R_{i,j}, L_{i,j}\right)}{\max\left(R_{i,j}, L_{i,j}\right)} = \frac{1 - abs\left(BSI_{i,j}\right)}{1 + abs\left(BSI_{i,j}\right)} \tag{3.11}$$

P_{min} denotes the relative power of the channel with the least power $(\min\left(R_{i,j}, L_{i,j}\right))$ compared to the power in channel with the most power $(\max\left(R_{i,j}, L_{i,j}\right))$. The relative power of the channel with the most power is by definition 1, since it is compared with itself.
The total power found within the two channels is defined as,

$$P_{max} = 1 + P_{min} \tag{3.12}$$

The value of P_{max} lies between 2 and 1, if $P_{max} = 2$ no asymmetry is precent since P_{min}must be 1.
If $P_{max} = 1$, maximal asymmetry is present meaning that no power was found in one of the channels.

To calculate the deviation from the healthy value, the reference matrix from the *tBSI* is used. However, the power in the reference matrix is based on healthy individuals. Any asymmetry will always lead to deviation from the reference even if the unaffected hemisphere is completely normal. To counter this phenomenon, the reference power is defined as

$$Ref_{asym} = P_{max} \times S_{ref\ i,j,a} \tag{3.13}$$

We now have a reference power that corrects for the asymmetry. Using a similar method as used in the *tBSI* the correction factor is defined as

$$C = 1 - \left|\frac{\left(R_{i,j}, + L_{i,j}\right) - Ref_{asym}}{\left(R_{i,j}, + L_{i,j}\right) + Ref_{asym}}\right| \tag{3.14}$$

creating a correction factor that will approach zero when the total power deviates from the healthy reference value. Implementing this correction factor in the *BSI* gives us

$$BSI_{corr} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{M}\sum_{j=1}^{M}\left(\frac{R_{i,j}-L_{i,j}}{R_{i,j}+L_{i,j}}\right)\times C\right| \tag{3.15}$$

After implementing both the *BSI* and the BSI_{corr} *the* performance was evaluated using the AUCROC. As seen in table 3, the correction factor worsens the performance. It can be seen that the mean prediction values of the BSI_{corr} are slower resulting in a smaller difference between the average prediction of the patient and control group.

Table 3. 4 Performance of the BSI and the BSI_{corr} evaluated at the using the AUCROC

	Mean prediction ± std (patients)	Mean prediction ± std (controls)	AUCROC	Sensitivity at 99% specificity	Sensitivity at 99% specificity
BSI	0.346 ± 0.118	0.205 ± 0.040	**0.95**	**72%**	**77%**
BSI_{corr}	0.201 ± 0.064	0.114 ± 0.038	0.91	61%	63%

Visualization of the algorithms

Both the *BSI* and *tBSI* with 4 frequency bins outperform all models in their class. To give an impression of their performance, the *BSI* and *tBSI* are visualized with the corresponding EEG.

As seen in figure 3.2, the bipolar derivations on the left hemisphere (odd numbers) do project EEG with higher frequencies and amplitude compared to the right hemisphere. This is most noticeable from 00:29:44 to 00:30:00 where bursts of faster activity are present in Fp1-F7, F7-T3 and T3-T5. This increased activity is lacking in Fp2-F8, F8-T4 and T4-T6. The activity of the left hemisphere gradually increases during this time frame, while the activity of the right hemisphere does not change much. This results in a greater asymmetry over time, which is also indicated by the *BSI*.

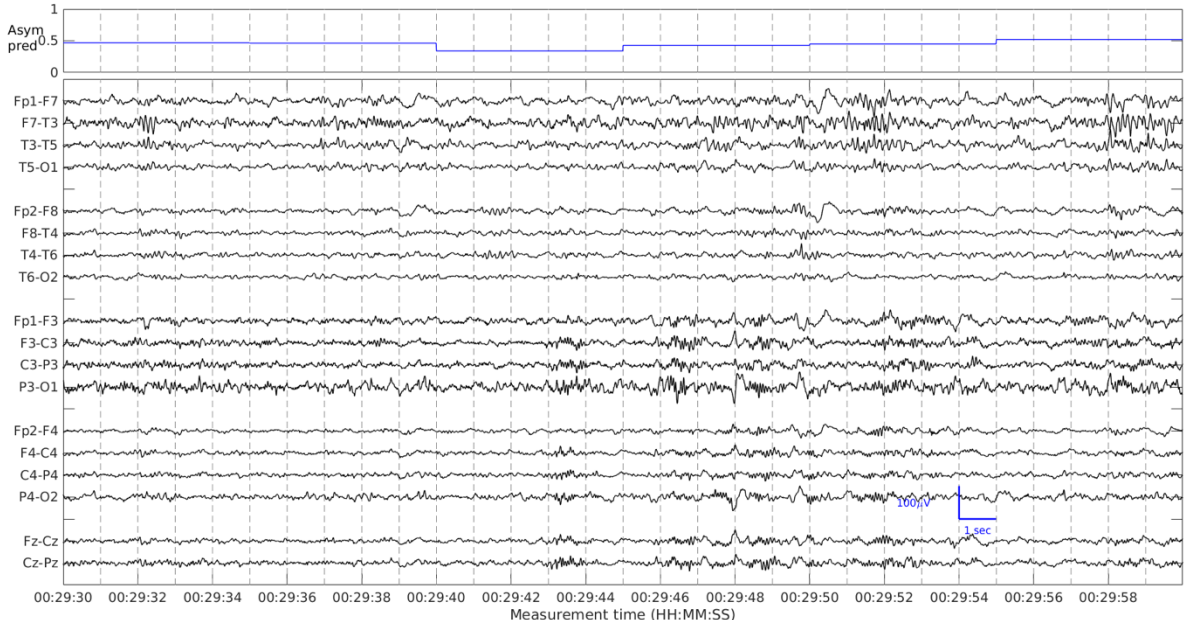

Figure 3. 2 The BSI as shown in the upper panel, and the corresponding EEG in the lower panel. This EEG does show asymmetry with the loss of fast activities at the right hemisphere. When the asymmetry worsens, the BSI does increase as can bes een from 00:29:40 to 00:30:00.

Figure 3.3 shows a EEG where background slowing is present. It can be seen, that the fast activities in all channels fade away in this time frame, where the *tBSI* increases in response to the slowing EEG.

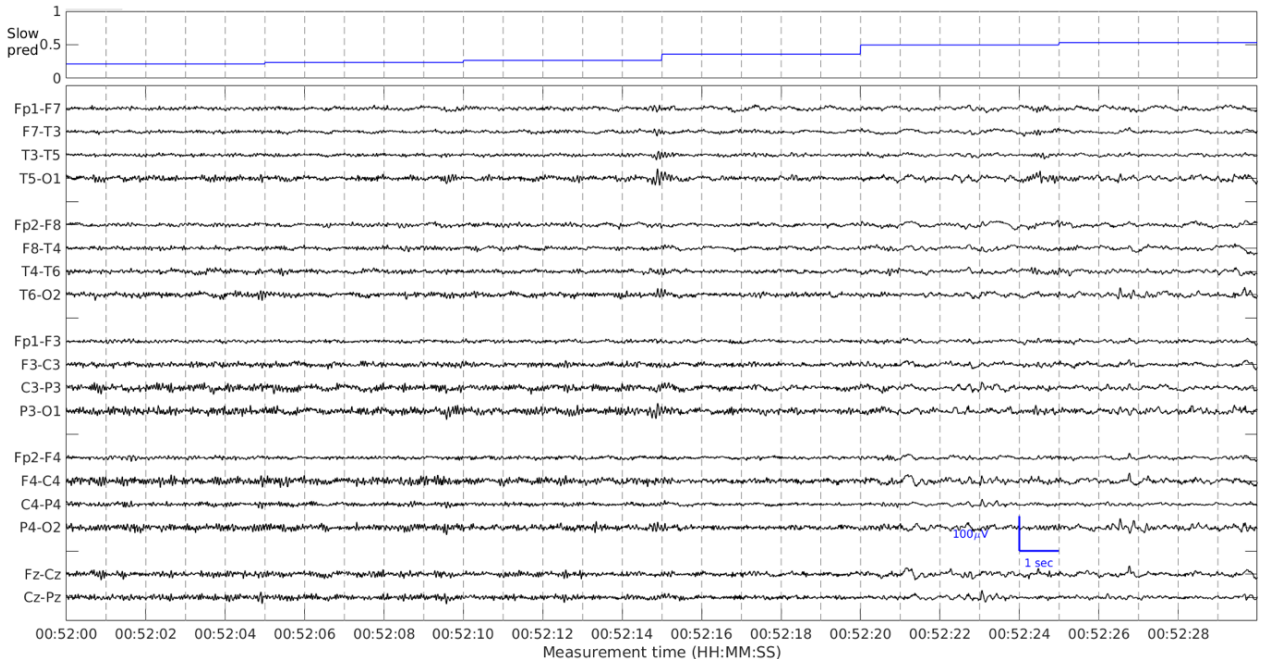

Figure 3. 3 The tBSI shown in the upper panel, shows a steady increase while the EEG slows. It can be seen that the high frequencies, wich are more dominant on the right, fade away. From 00:52:21 and on, the slow wave activity becomes more prominent. Both fenomenon lead to a slower EEG, which is also indictadet by the tBSI.

Discussion

After implementing the *BSI, r-BSI, tBSI and r-tBSI* and selecting the best performing models for further experiments, the *BSI* and *tBSI* using 4 frequency bins, corresponding to the delta, theta, alpha and theta range of the EEG, reaches the highest AUCROC with a value of 0.95 and 0.87 respectively. The sensitivity at 99% and 95% specificity shows that using 4 frequency bins, chosen based on physiological knowledge, improves the performance by elevating the sensitivity at 99% specificity for *BSI* and *tBSI* with respectively 4% and 9% and the sensitivity at 95% specificity with respectively 1% and 6%. For the (*r-*)*tBSI* another contribution was made by showing that our reference matrix created by our DNN is able to function as a reference for predicting (near) continuous slowing. Making it not only useful during carotid endarterectomy but also as a general (near) continuous slowing detection algorithm.

Lodder et al. proposed an automatic detection algorithm for diffuse slow-wave activity with a sensitivity of 78% at a specificity of 98% [89]. Although the performance is quite similar, Lodder et al. does outperform the *tBSI*. The algorithm from Lodder et al. does need an extra input feature since this input allows only segments from the eyes closed state. This extra information may not always be available resulting in a less applicable model.

A major weakness in this study is the lack of (epoch wise) labels, resulting in a free text search to create the patients and control groups. The labels, given to all epochs in the patient, are based on the patients report, inevitable resulting in misclassified epoch. Especially since near continuous slowing and asymmetry were accepted in the patient groups. Epoch wise labeling will enable us to do a more precise study which could lead to further optimalisation of this model. However, our results do pave a way for a follow-up study regarding the detection of intermitted slowing and asymmetry.

The strength of our study is the tiered and broad validation of the , the *BSI/tBSI* and the *r-BSI/r-tBSI.* Resulting in the surprising result that the original *BSI* and *tBSI* outperform the *r-BSI* and *r-tBSI,* which we belive lies in the mathematical basis of the formula. The difference in the approach between the *BSI/tBSI* and the *r-BSI/r-tBSI* which causes the biggest change in the outputted value is that the *r-BSI/r-tBSI* is taking the squared power of a segment, as can be seen in equation 3.3 and 3.4. Taking the squared power does increase the sensitivity of the algorithm resulting in the bigger difference in the mean value between the patient and control group. However, this also effects the standard deviation resulting in a slightly worse performance. The standard deviation might be reduced when epoch wise labeling is present, eradicating the incorrectly labeled epochs and boosting the performance.

The BSI_{corr} does, as expected, predict lower values compared to the *BSI.* However, a bigger drop in the outputted value is present patient group compared to the control group, resulting in a smaller difference between the patient and control group. The smaller difference makes is harder to distinguish between the groups leading to a performance drop. The difference in the outputted value from the BSI_{corr} indicates that the correction factor in the patient group was overall lower, and therefore correcting more, compared to the control group. This may be caused by the overlap of slowing patients in the asymmetry group. It was chosen to include patient with asymmetry and slowing into the asymmetry group since simultaneous occurrence of both conditions does exist and should be taken into account while optimizing the algorithm.

In conclusion, we validated the earlier presented *BSI* and *r-BSI.* Using a selective approach we improved the *BSI* by reducing the frequency bins to follow the four physiological defined frequency bands of the EEG leading to an AUCROC of 0.95. Secondly, were able to generate a smooth general reference matrix using a deep neural net. The reference matrix was used in the validation of the the *r-tBSI* and *tBSI* for detecting continuous slowing. Using the same selective improvement approach, the AUCROC of the *tBSI* was increased to 0.88. Both the *BSI* and the *tBSI* do yield promising results on continuous slowing and asymmetry, future work is needed to evaluate the performance of intermitted slowing and asymmetry.

General conclusion

Based on the previous, we could state that our work is a valuable contribution towards the automated analysis of the EEG. We validated multiple asymmetry detection algorithms, enhanced a slowing detection algorithm, reduced the FP/h in a state-of-the-art IED detector and we paved the way for GAN generated EEG segments.

To answer our main research question, we first need to answer our sub questions.
In chapter one, where we address the false positive detections from the IED detector, we answered the question:

> *To what extend will using a hard example mining method, reduce the false positive prediction of the interictal epileptiform discharge detector?*

We showed an increase in the $AUCROC_{ajust}$ and $AUCPRC_{ajust}$ of at least 0.9985 and 0.9983 respectively. Most importantly, we showed a decrease up to 70% in the false detection rate, resulting in a FP/h of 15 at a $95\%_{5/8}$ sensitivity. Therefore, clear evidence is given that hard example mining does significantly improve the results and should be incorporated into the standard training procedure of similar models.

In chapter two, we tried to improve SpikeNet even further by using generated EEG segments in the training, therefore answering the following question:

> *To what extend can generated EEG segments increase the performance of the interictal epileptiform discharge detector?*

After overcoming multiple pitfalls regarding stability and labeling generated IED segments, we were able to improve the overall $AUCROC_{ajust}$ to 0.9986, however small decreases were found in the $AURPRC_{ajust}$. Monitoring the outputs closely, it can be found that the SpikeNet performance only increased when using a threshold of 0.28 and lower. This results in the increased FP/h of 18.3 at $95\%_{5/8}$ sensitivity. Since a high sensitivity is required for this task, we conclude that adding the generated IED's does not improve the model performance.

In conclusion, we showed promising results regarding the generation of IED's. Unfortunately, the generated data is too noisy in absolute terms but also too noisy regarding the labeling assignment, making it, at this stage, impossible to incorporated the generated data for training purposes. Leaving augmentation methods that heavily rely on the input data the preferred choice for enhancing SpikeNet.

Recommendations

Future steps for SpikeNet

To improve the performance of SpikeNet, it is necessary to further reduce the false positive rates while maintaining high sensitivities. During the evaluating the false positive predictions it is found that most false positive predictions are caused by artifacts. The artifacts that are not rejected by the background rejection algorithm are mostly difficult to reject using a rule-based algorithm. However, one artifact does seem to be a candidate for addition to the rule-based algorithm. The artifact that could be ruled out based fairly simple features, is the artifact that is created when the measurement starts or ends. These artifacts occur per definition adjacent to a flatline. Flatlines are already to detect by the rejection algorithm which makes it possible to extend the rejection, so it includes the artifact.

Secondly, some benign variants of uncertain significance, such as vertex waves and sleep spindles, are related to sleep depth. Incorporating sleep depth as an input feature could potentially provide useful information. Even though this yield some predictive power, this is most likely not able to drastically improves the performance. Creating a multi-task model which predicts IED's and also benign variants is likely to improve the results [90]. Since the model is forced to learn both the features of the IED's and the benign variants, features that could distinguish between the latter two must be learned to achieve good performance in both tasks. The power of the multi-task model lies within the shared features of both tasks, where learning features for one task could additionally benefit another task since the features per task do partly overlap.

I would recommend starting with the expanding of the model with one benign variant that has a clearly distinguishable morphology, for example sleep spindles. Ahmed et al. showed a 93.7% accuracy for his automatic spindle detection making this a good candidate for an unsupervised sleep spindle mining algorithm [91]. Subsequently, the mined EEG can be used to train the multi-task model. The spindle detection could also be used as a stand-alone feature, if the multi-task model does not reach similar performance as the spindle detection from Ahmed et al.

Future steps for generating IED segments

The methods used for generating IED's are barely scratching the surface of all possibilities. In this paragraph we purpose some architectural and model specific changes that could potentially lead to an increased performance.
First of all, we recommend using residual blocks instead of the convolutional blocks that are currently used in the generator and discriminator. Residual blocks, firstly described by He et al., is an architecture to ease the training of very deep neural networks[92]. The residual connections, which function as shortcuts to the model, are enabling the training of deeper nets. In our study, it is found that increasing the depth of the architectures in the GAN is accompanied by increased instability. Reducing that instability by adding residual connections might enable us to learn more features without compromising on the stability. Increased features could lead to a better adaptation to the target domain in both the morphology as well as the variability of the generated EEG.

Secondly, we chose to only use IED's with the label 8/8 as input domain. Using a subset of IED's will not give a correct representation of the input domain. A Conditional GAN give you the opportunity to use all data from the target domain while enabling you to control which IED's you want to generate[93]. This is achieved by incorporating conditions as an additional input for both the generator as well as the discriminator. In our case we could incorporate the class labels as condition.

If the discriminator is able to learn the relation between the added labels and the IED's, it will penalize the generator if it generates IED's that do not correspond with the given label.
An additional advantage of this approach is that it will ease the implementation of the IED's in the SpikeNet dataset since the generator is able to generate the IED's by label.

Lastly, we are suggesting a modification of the loss function. Luo et al. introduced a new loss function special for EEG synthesis, the so called spatio-temporal-frequency loss, which is an addition to the current loss as described in equation 2.3 to 2.5 [67]. This spatio-temporal-frequency loss is based on spatial, temporal and frequency features of the EEG and penalizes the GAN to whenever it deviates from the optimal features. Incorporating this loss should encourage the model into generating EEG segments with an more realistic spatial, temporal and frequency features.

Future steps for the slowing and asymmetry detection

We showed the potential of both the *BSI* and the *tBSI* by demonstrating excellent performance for both the *BSI* and the *tBSI* using patients with (near) continuous slowing and asymmetry. To make the algorithms applicable for clinical usage, they should also be able to detect intermitted slowing and asymmetry. To calibrate and extend the model for intermitted patterns, epoch wise labeling is preferred. The *tBSI* and *BSI* can be used to identify potential epochs that contain intermitted slowing and asymmetry accordingly. Finally, the proposed epochs should be rated by multiple raters to identify the if the proposed epochs indeed contain slowing or asymmetry.

After implementing the correction factor in the *BSI* the performance dropped. A plausible explanation is the overlap of patients with generalized slowing and asymmetry. The correction factor is built in such way that it functions correctly if one hemisphere is unaffected. The correction factor will output lower values, and therefore correct more, when in addition to the asymmetry, generalized slowing is present. To overcome this burden, the *tBSI* should be incorporated into the correction factor.

References

[1] V. L. Feigin *et al.*, "Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016," *Lancet Neurol.*, vol. 18, no. 5, pp. 459–480, May 2019, doi: 10.1016/S1474-4422(18)30499-X.

[2] T. Dua, World Federation of Neurology, World Health Organization, and World Health Organization, Eds., *Atlas: country resources for neurological disorders 2004: results of a collaborative study of the World Health Organization and the World Federation of Neurology*. Geneva: Programme for Neurological Diseases and Neuroscience, Department of Mental Health and Substance Abuse, World Health Organization, 2004.

[3] P. Miranda, C. D Cox, M. Alexander, S. Danev, and J. RT Lakey, "Overview of current diagnostic, prognostic, and therapeutic use of EEG and EEG-based markers of cognition, mental, and brain health," *Integr. Mol. Med.*, vol. 6, no. 5, 2019, doi: 10.15761/IMM.1000378.

[4] J. Jing *et al.*, "Development of Expert-Level Automated Detection of Epileptiform Discharges During Electroencephalogram Interpretation," *JAMA Neurol.*, vol. 77, no. 1, p. 103, Jan. 2020, doi: 10.1001/jamaneurol.2019.3485.

[5] J. Jing *et al.*, "Interrater Reliability of Experts in Identifying Interictal Epileptiform Discharges in Electroencephalograms," *JAMA Neurol.*, vol. 77, no. 1, p. 49, Jan. 2020, doi: 10.1001/jamaneurol.2019.3531.

[6] M. C. Tjepkema-Cloostermans, R. C. V. de Carvalho, and M. J. A. M. van Putten, "Deep learning for detection of focal epileptiform discharges from scalp EEG recordings," *Clin. Neurophysiol.*, vol. 129, no. 10, pp. 2191–2196, Oct. 2018, doi: 10.1016/j.clinph.2018.06.024.

[7] A. Ulate-Campos, F. Coughlin, M. Gaínza-Lein, I. S. Fernández, P. L. Pearl, and T. Loddenkemper, "Automated seizure detection systems and their effectiveness for each type of seizure," *Seizure*, vol. 40, pp. 88–101, Aug. 2016, doi: 10.1016/j.seizure.2016.06.008.

[8] P. E. Rapp *et al.*, "Traumatic Brain Injury Detection Using Electrophysiological Methods," *Front. Hum. Neurosci.*, vol. 9, Feb. 2015, doi: 10.3389/fnhum.2015.00011.

[9] A. Tolonen *et al.*, "Quantitative EEG Parameters for Prediction of Outcome in Severe Traumatic Brain Injury: Development Study," *Clin. EEG Neurosci.*, vol. 49, no. 4, pp. 248–257, Nov. 2017, doi: 10.1177/1550059417742232.

[10] M. Rots, M. J. A. M. van Putten, C. W. E. Hoedemaekers, and J. Horn, "Continuous EEG Monitoring for Early Detection of Delayed Cerebral Ischemia in Subarachnoid Hemorrhage: A Pilot Study," *Neurocrit. Care*, vol. 24, no. 2, p. 207—216, Apr. 2016, doi: 10.1007/s12028-015-0205-y.

[11] S. Gollwitzer *et al.*, "Quantitative EEG After Subarachnoid Hemorrhage Predicts Long-Term Functional Outcome," *J. Clin. Neurophysiol. Off. Publ. Am. Electroencephalogr. Soc.*, vol. 36, no. 1, p. 25—31, Jan. 2019, doi: 10.1097/wnp.0000000000000537.

[12] M. Symmonds *et al.*, "Ion channels in EEG: isolating channel dysfunction in NMDA receptor antibody encephalitis," *Brain*, vol. 141, no. 6, pp. 1691–1702, Jun. 2018, doi: 10.1093/brain/awy107.

[13] Y. Wu *et al.*, "Viral encephalitis in quantitative EEG," *J. Integr. Neurosci.*, vol. 17, no. 3–4, p. 493—501, 2018, doi: 10.3233/jin-180084.

[14] F. N. Karameh and M. A. Dahleh, "Automated classification of EEG signals in brain tumor diagnostics," in *Proceedings of the 2000 American Control Conference. ACC (IEEE Cat. No.00CH36334)*, Jun. 2000, vol. 6, pp. 4169–4173 vol.6, doi: 10.1109/ACC.2000.877006.

[15] V. S. Selvam and S. S. Devi, "Analysis of Spectral Features of EEG signal in Brain Tumor Condition," *Meas. Sci. Rev.*, vol. 15, no. 4, pp. 219–225, Aug. 2015, doi: 10.1515/msr-2015-0030.

[16] L. C. Weeke, A. Vilan, M. C. Toet, I. C. van Haastert, L. S. de Vries, and F. Groenendaal, "A Comparison of the Thompson Encephalopathy Score and Amplitude-Integrated Electroencephalography in Infants with Perinatal Asphyxia and Therapeutic Hypothermia," *Neonatology*, vol. 112, no. 1, pp. 24–29, 2017, doi: 10.1159/000455819.

[17] D. M. Murray, C. M. OConnor, C. A. Ryan, I. Korotchikova, and G. B. Boylan, "Early EEG Grade and Outcome at 5 Years After Mild Neonatal Hypoxic Ischemic Encephalopathy," *PEDIATRICS*, vol. 138, no. 4, pp. e20160659–e20160659, Oct. 2016, doi: 10.1542/peds.2016-0659.

[18] V. Kavcic, B. Zalar, and B. Giordani, "The relationship between baseline EEG spectra power and memory performance in older African Americans endorsing cognitive concerns in a community

setting," *Int. J. Psychophysiol.*, vol. 109, pp. 116–123, Nov. 2016, doi: 10.1016/j.ijpsycho.2016.09.001.
[19] A. D. Kalechstein, R. De La Garza, T. F. Newton, M. F. Green, I. A. Cook, and A. F. Leuchter, "Quantitative EEG Abnormalities are Associated With Memory Impairment in Recently Abstinent Methamphetamine-Dependent Individuals," *J. Neuropsychiatry*, vol. 21, no. 3, pp. 254–258, Aug. 2009, doi: 10.1176/appi.neuropsych.21.3.254.
[20] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *J. Am. Med. Inform. Assoc.*, vol. 25, no. 12, pp. 1643–1650, Dec. 2018, doi: 10.1093/jamia/ocy131.
[21] H. Sun *et al.*, "Large-Scale Automated Sleep Staging," *Sleep*, vol. 40, no. 10, Oct. 2017, doi: 10.1093/sleep/zsx139.
[22] V. Latreille, M. Gaubert, J. Dubé, J.-M. Lina, J.-F. Gagnon, and J. Carrier, "Age-related cortical signatures of human sleep electroencephalography," *Neurobiol. Aging*, vol. 76, pp. 106–114, Apr. 2019, doi: 10.1016/j.neurobiolaging.2018.12.012.
[23] J. B. Stephansen *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nat. Commun.*, vol. 9, no. 1, Dec. 2018, doi: 10.1038/s41467-018-07229-3.
[24] E. Westhall *et al.*, "Interrater variability of EEG interpretation in comatose cardiac arrest patients," *Clin. Neurophysiol.*, vol. 126, no. 12, pp. 2397–2404, Dec. 2015, doi: 10.1016/j.clinph.2015.03.017.
[25] M. M. Admiraal *et al.*, "Electroencephalographic reactivity as predictor of neurological outcome in postanoxic coma: A multicenter prospective cohort study," *Ann. Neurol.*, vol. 86, no. 1, pp. 17–27, Jul. 2019, doi: 10.1002/ana.25507.
[26] J. Brogger, T. Eichele, E. Aanestad, H. Olberg, I. Hjelland, and H. Aurlien, "Visual EEG reviewing times with SCORE EEG," *Clin. Neurophysiol. Pract.*, vol. 3, pp. 59–64, 2018, doi: 10.1016/j.cnp.2018.03.002.
[27] E. K. St. Louis, L. C. Frey, J. W. Britton, and American Epilepsy Society, *Electroencephalography (EEG): an introductory text and atlas of normal and abnormal findings in adults, children, and infants*. 2016.
[28] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *ArXiv171204621 Cs*, Dec. 2017, Accessed: Oct. 11, 2020. [Online]. Available: http://arxiv.org/abs/1712.04621.
[29] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard, "Adaptive data augmentation for image classification," in *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 3688–3692, doi: 10.1109/ICIP.2016.7533048.
[30] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018, doi: 10.1016/j.neucom.2018.09.013.
[31] F. Wang, S. Zhong, J. Peng, J. Jiang, and Y. Liu, "Data Augmentation for EEG-Based Emotion Recognition with Deep Convolutional Neural Networks," in *MultiMedia Modeling*, Cham, 2018, pp. 82–93.
[32] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, Swinoujście, May 2018, pp. 117–122, doi: 10.1109/IIPHDW.2018.8388338.
[33] A. Shrivastava, A. Gupta, and R. Girshick, "Training Region-Based Object Detectors with Online Hard Example Mining," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 761–769, doi: 10.1109/CVPR.2016.89.
[34] T. Jaakkola and D. Haussler, "Exploiting Generative Models in Discriminative Classifiers," p. 7.
[35] N. K. N. Aznan, A. Atapour-Abarghouei, S. Bonner, J. Connolly, N. A. Moubayed, and T. Breckon, "Simulating Brain Signals: Creating Synthetic EEG Data via Neural-Based Generative Models for Improved SSVEP Classification," *2019 Int. Jt. Conf. Neural Netw. IJCNN*, pp. 1–8, Jul. 2019, doi: 10.1109/IJCNN.2019.8852227.
[36] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," p. 10.

[37] A. C. Quiros, R. Murray-Smith, and K. Yuan, “PathologyGAN: Learning deep representations of cancer tissue,” p. 29.
[38] “EEG (electroencephalogram) - Mayo Clinic.” https://www.mayoclinic.org/tests-procedures/eeg/about/pac-20393875 (accessed Feb. 07, 2020).
[39] S. Seidel, E. Pablik, S. Aull-Watschinger, B. Seidl, and E. Pataraia, “Incidental epileptiform discharges in patients of a tertiary centre,” *Clin. Neurophysiol.*, vol. 127, no. 1, pp. 102–107, Jan. 2016, doi: 10.1016/j.clinph.2015.02.056.
[40] C. A. van Donselaar, R.-J. Schimsheimer, A. T. Geerts, and A. C. Declerck, “Value of the Electroencephalogram in Adult Patients With Untreated Idiopathic First Seizures,” *Arch. Neurol.*, vol. 49, no. 3, pp. 231–237, Mar. 1992, doi: 10.1001/archneur.1992.00530270045017.
[41] N. B. Fountain and J. M. Freeman, “EEG Is an Essential Clinical Tool: Pro and Con,” *Epilepsia*, vol. 47, no. s1, pp. 23–25, Oct. 2006, doi: 10.1111/j.1528-1167.2006.00655.x.
[42] S. Noachtar and J. Rémi, “The role of EEG in epilepsy: A critical review,” *Epilepsy Behav.*, vol. 15, no. 1, pp. 22–33, May 2009, doi: 10.1016/j.yebeh.2009.02.035.
[43] R. S. Rosenberg and S. Van Hout, “The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring,” *J. Clin. Sleep Med.*, Jan. 2013, doi: 10.5664/jcsm.2350.
[44] A. C. Grant *et al.*, “EEG interpretation reliability and interpreter confidence: A large single-center study,” *Epilepsy Behav.*, vol. 32, pp. 102–107, Mar. 2014, doi: 10.1016/j.yebeh.2014.01.011.
[45] A. Y. Hannun *et al.*, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nat. Med.*, vol. 25, no. 1, pp. 65–69, Jan. 2019, doi: 10.1038/s41591-018-0268-3.
[46] Z. Cui, X. Zheng, X. Shao, and L. Cui, “Automatic Sleep Stage Classification Based on Convolutional Neural Network and Fine-Grained Segments,” *Complexity*, vol. 2018, pp. 1–13, Oct. 2018, doi: 10.1155/2018/9248410.
[47] M. L. Scheuer, A. Bagic, and S. B. Wilson, “Spike detection: Inter-reader agreement and a statistical Turing test on a large data set,” *Clin. Neurophysiol.*, vol. 128, no. 1, pp. 243–250, Jan. 2017, doi: 10.1016/j.clinph.2016.11.005.
[48] I. Drury and A. Beydoun, “Pitfalls of EEG interpretation in epilepsy,” *Neurol. Clin.*, vol. 11, no. 4, p. 857—881, Nov. 1993, doi: 10.1016/s0733-8619(18)30128-2.
[49] S. R. Benbadis and K. Lin, “Errors in EEG Interpretation and Misdiagnosis of Epilepsy,” *Eur. Neurol.*, vol. 59, no. 5, pp. 267–271, 2008, doi: 10.1159/000115641.
[50] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object Detection with Discriminatively Trained Part Based Models,” p. 20.
[51] K. Sung and T. Poggio, “Example Based Learning for View-Based Human Face Detection,” *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 20, pp. 39–51, 1998, doi: 10.1109/34.655648.
[52] R. T. Schirrmeister *et al.*, “Deep learning with convolutional neural networks for EEG decoding and visualization: Convolutional Neural Networks in EEG Analysis,” *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017, doi: 10.1002/hbm.23730.
[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[54] W. O. Tatum, D. Schmidt, W. O. Tatum, and S. Schachter, “Mistaking EEG Changes for Epilepsy,” in *Common Pitfalls in Epilepsy: Case-Based Learning*, Cambridge University Press, 2018, pp. 25–42.
[55] G. Olmos G de Alba, M. I. Fraire-Martínez, and R. Valenzuela-Romero, “Clinical correlation of hypnagogic hypersynchrony during sleep in normal children and those with learning disability.,” *Rev. Neurol.*, vol. 36, no. 8, pp. 720–723, Apr. 2003.
[56] R. B. Berry and M. H. Wagner, “Introduction,” in *Sleep Medicine Pearls*, Elsevier, 2015, pp. 10–14.
[57] J. M. Stern and J. Engel, “Atlas of EEG Patterns,” p. 467, 2015.
[58] M. S. A. Batista, C. F. Coelho, M. M. D. Lima, and D. F. Silva, “Wicket spikes: a case-control study of a benign eletroencephalografic variant pattern,” *Arq. Neuropsiquiatr.*, vol. 57, no. 3A, pp. 561–565, Sep. 1999, doi: 10.1590/S0004-282X1999000400004.
[59] R. Flink *et al.*, “Guidelines for the use of EEG methodology in the diagnosis of epilepsy:

International League Against Epilepsy: Commission Report Commission on European Affairs: Subcommission on European Guidelines," *Acta Neurol. Scand.*, vol. 106, no. 1, pp. 1–7, Jul. 2002, doi: 10.1034/j.1600-0404.2002.01361.x.
[60] J. Brownlee, "Generative Adversarial Networks with Python," p. 654.
[61] A. Y. Ng and M. I. Jordan, "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes," p. 8.
[62] I. Goodfellow *et al.*, "Generative Adversarial Nets," p. 9.
[63] H. Huang, "IntroVAE: Introspective Variational Autoencoders for Photographic Image Synthesis," p. 12.
[64] P. Costa *et al.*, "Towards Adversarial Retinal Image Synthesis," *ArXiv170108974 Cs Stat*, Jan. 2017, Accessed: Oct. 12, 2020. [Online]. Available: http://arxiv.org/abs/1701.08974.
[65] S. M. Abdelfattah, G. M. Abdelrahman, and M. Wang, "Augmenting The Size of EEG datasets Using Generative Adversarial Networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Jul. 2018, pp. 1–6, doi: 10.1109/IJCNN.2018.8489727.
[66] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, "EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals," *ArXiv180601875 Cs Eess Q-Bio Stat*, Jun. 2018, Accessed: Oct. 12, 2020. [Online]. Available: http://arxiv.org/abs/1806.01875.
[67] T. Luo, Y. Fan, L. Chen, G. Guo, and C. Zhou, "EEG Signal Reconstruction Using a Generative Adversarial Network With Wasserstein Distance and Temporal-Spatial-Frequency Loss," *Front. Neuroinformatics*, vol. 14, Apr. 2020, doi: 10.3389/fninf.2020.00015.
[68] I. Goodfellow, "NIPS 2016 Tutorial: Generative Adversarial Networks," *ArXiv170100160 Cs*, Apr. 2017, Accessed: Oct. 12, 2020. [Online]. Available: http://arxiv.org/abs/1701.00160.
[69] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs Created Equal? A Large-Scale Study," *ArXiv171110337 Cs Stat*, Oct. 2018, Accessed: Feb. 06, 2020. [Online]. Available: http://arxiv.org/abs/1711.10337.
[70] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved Training of Wasserstein GANs," p. 11.
[71] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," *ArXiv160603498 Cs*, Jun. 2016, Accessed: Oct. 13, 2020. [Online]. Available: http://arxiv.org/abs/1606.03498.
[72] Z. Zhang, "Improved Adam Optimizer for Deep Neural Networks," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, Banff, AB, Canada, Jun. 2018, pp. 1–2, doi: 10.1109/IWQoS.2018.8624183.
[73] S. Ruder, "An overview of gradient descent optimization algorithms," *ArXiv160904747 Cs*, Jun. 2017, Accessed: Oct. 13, 2020. [Online]. Available: http://arxiv.org/abs/1609.04747.
[74] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[75] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *ArXiv161107004 Cs*, Nov. 2018, Accessed: Oct. 13, 2020. [Online]. Available: http://arxiv.org/abs/1611.07004.
[76] J. Ding, X. Ren, R. Luo, and X. Sun, "An Adaptive and Momental Bound Method for Stochastic Learning," *ArXiv191012249 Cs Stat*, Oct. 2019, Accessed: Oct. 13, 2020. [Online]. Available: http://arxiv.org/abs/1910.12249.
[77] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *ArXiv151106434 Cs*, Jan. 2016, Accessed: Oct. 13, 2020. [Online]. Available: http://arxiv.org/abs/1511.06434.
[78] P. W. Kaplan and A. O. Rossetti, "EEG Patterns and Imaging Correlations in Encephalopathy: Encephalopathy Part II," *J. Clin. Neurophysiol.*, vol. 28, no. 3, 2011, [Online]. Available: https://journals.lww.com/clinicalneurophys/Fulltext/2011/06000/EEG_Patterns_and_Imaging_Correlations_in.1.aspx.
[79] C. Petit, R. Bezemer, and L. Atallah, "A review of recent advances in data analytics for post-operative patient deterioration detection," *J. Clin. Monit. Comput.*, vol. 32, no. 3, pp. 391–402, Jun. 2018, doi: 10.1007/s10877-017-0054-7.
[80] J. Hofmeijer, M. C. Tjepkema-Cloostermans, and M. J. A. M. van Putten, "Burst-suppression with identical bursts: A distinct EEG pattern with poor outcome in postanoxic coma," *Clin. Neurophysiol.*, vol. 125, no. 5, pp. 947–954, May 2014, doi: 10.1016/j.clinph.2013.10.017.

[81] C. Bentes *et al.*, “Quantitative EEG and functional outcome following acute ischemic stroke,” *Clin. Neurophysiol.*, vol. 129, no. 8, pp. 1680–1687, Aug. 2018, doi: 10.1016/j.clinph.2018.05.021.
[82] C. Bentes *et al.*, “Early EEG predicts poststroke epilepsy,” *Epilepsia Open*, vol. 3, no. 2, pp. 203–212, Jun. 2018, doi: 10.1002/epi4.12103.
[83] A. Boyer, S. Ramdani, H. Duffau, B. Poulin-Charronnat, D. Guiraud, and F. Bonnetblanc, “Alterations of EEG rhythms during motor preparation following awake brain surgery,” *Brain Cogn.*, vol. 125, pp. 45–52, Aug. 2018, doi: 10.1016/j.bandc.2018.05.010.
[84] M. Van Putten, “Extended BSI for continuous EEG monitoring in carotid endarterectomy,” *Clin. Neurophysiol.*, vol. 117, no. 12, pp. 2661–2666, Dec. 2006, doi: 10.1016/j.clinph.2006.08.007.
[85] M. J. A. M. van Putten, J. M. Peters, S. M. Mulder, J. A. M. de Haas, C. M.A. Bruijninckx, and D. L. J. Tavy, “A brain symmetry index (BSI) for online EEG monitoring in carotid endarterectomy,” *Clin. Neurophysiol.*, vol. 115, no. 5, pp. 1189–1194, May 2004, doi: 10.1016/j.clinph.2003.12.002.
[86] M. J. A. M. van Putten, “The revised brain symmetry index,” *Clin. Neurophysiol.*, vol. 118, no. 11, pp. 2362–2367, Nov. 2007, doi: 10.1016/j.clinph.2007.07.019.
[87] B. Scally, M. R. Burke, D. Bunce, and J.-F. Delvenne, “Resting-state EEG power and connectivity are associated with alpha peak frequency slowing in healthy aging,” *Neurobiol. Aging*, vol. 71, pp. 149–155, Nov. 2018, doi: 10.1016/j.neurobiolaging.2018.07.004.
[88] J. Carrier, S. Land, D. J. Buysse, D. J. Kupfer, and T. H. Monk, “The effects of age and gender on sleep EEG power spectral density in the middle years of life (ages 20-60 years old),” *Psychophysiology*, vol. 38, no. 2, pp. 232–242, Mar. 2001, doi: 10.1111/1469-8986.3820232.
[89] S. S. Lodder and M. J. A. M. van Putten, “Quantification of the adult EEG background pattern,” *Clin. Neurophysiol.*, vol. 124, no. 2, pp. 228–237, Feb. 2013, doi: 10.1016/j.clinph.2012.07.007.
[90] R. Collobert and J. Weston, “A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning,” p. 8.
[91] B. Ahmed, A. Redissi, and R. Tafreshi, “An automatic sleep spindle detector based on wavelets and the teager energy operator,” in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Sep. 2009, pp. 2596–2599, doi: 10.1109/IEMBS.2009.5335331.
[92] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *ArXiv151203385 Cs*, Dec. 2015, Accessed: Nov. 10, 2020. [Online]. Available: http://arxiv.org/abs/1512.03385.
[93] M. Mirza and S. Osindero, “Conditional Generative Adversarial Nets,” *ArXiv14111784 Cs Stat*, Nov. 2014, Accessed: Feb. 06, 2020. [Online]. Available: http://arxiv.org/abs/1411.1784.

Appendices

Appendix 1: Background detection algorithm

The background rejection algorithm that is used for the rejection artifacts from the EEG is an in-house build, rule-based algorithm built in MATLAB. The background rejection algorithm calculates features based upon 1 second epochs of non-overlapping EEG. The following features are calculated:

- average zero crossings
- max (EEG)
- $sum(P_{full})$
- max (P_{theta}/P_{gamma})
- mean (P_{theta}/P_{gamma})
- max (P_{theta})
- mean (P_{theta})
- max (P_{alpha}/P_{gamma})
- mean (P_{alpha}/P_{gamma})
- mean (P_{delta})
- mean (P_{theta}/P_{alpha})
- max (P_{beta}/P_{gamma})

Where P_f is denoting the average band power per channel for frequency range f. For each of the calculated values, a threshold value is present and if one of the calculated features exceeds its personal threshold, the 1 second EEG segment is rejected.

Appendix 2: Grad-CAM

Grad-CAM is a technique for making convolutional neural networks transparent by visualizing input regions on which the model based its prediction[53]. It can be seen in figure a. 2.1, that Grad-CAM is a stand-alone model which taps data from the original model to calculate the localization map.

To obtain the class-discriminative localization map, the gradient score of class c, y^c, is calculated with respect to the feature map activations A^k, and is denoted as $\frac{\partial y^c}{\partial A^k}$. To obtain the neuron importance weight α_k^c, the gradient score, with the size $i \times j$, is global-average-pooled and can be denoted as,

$$a_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \qquad (a.2.1)$$

with

$$Z = i \times j \qquad (a.2.2)$$

The localization heatmap is created by a weighted combination of feature maps which are passed through a ReLU, which can be denoted as,

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \qquad (a.2.3)$$

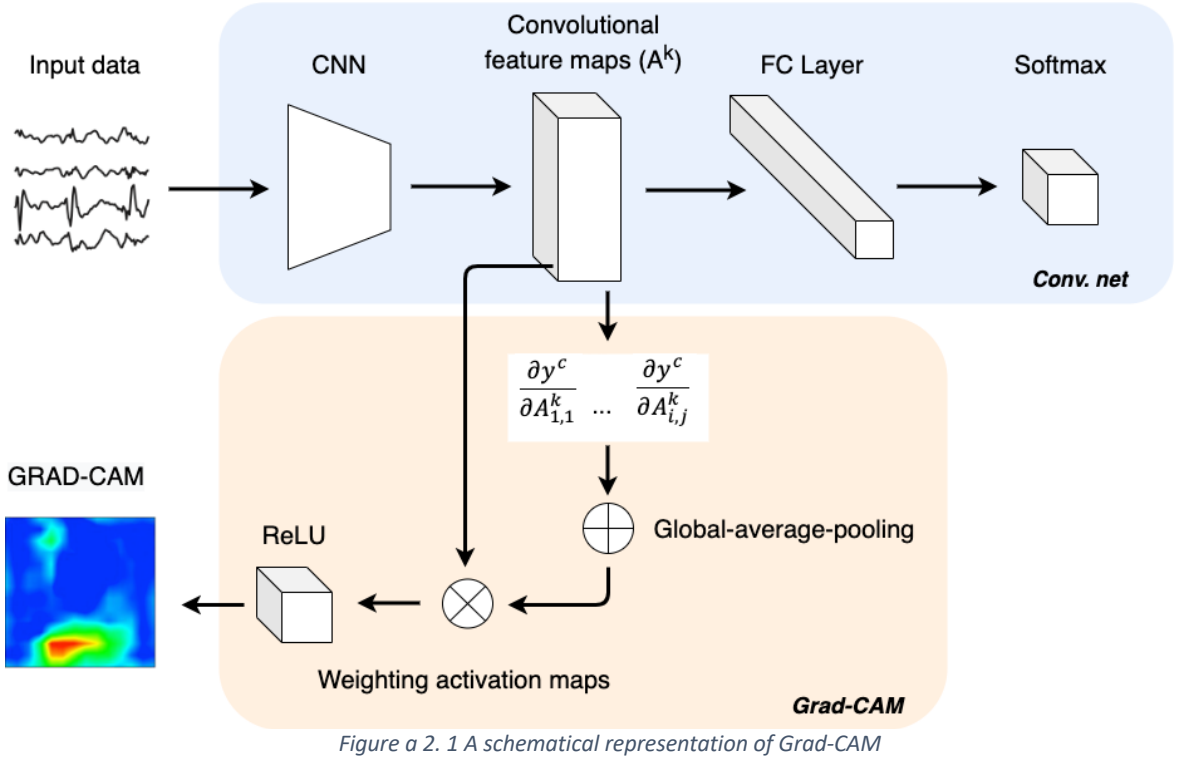

Figure a 2. 1 A schematical representation of Grad-CAM