REI505M Gervigreind

Neonatal seizure detection

Jóhannes Reykdal Einarsson
Stefán Árni Arnarsson
Sölvi Santos

November 22, 2024

REI505M Final Project
School of Engineering and Natural Sciences

Contents

Neonatal seizure detection

Introduction

A critical challenge in neonatal intensive care units (NICU) is seizure detection in newborns. This is due to the potentially devastating effects of prolonged seizures. These include permanent brain damage and severe disabilities. It is of the utmost importance to detect these seizures since it allows for the administration of anti-epileptic drugs to halt seizures. The largest hindrance regarding this matter is that neonatal seizures often go unnoticed. Only about one-third of them are clinically visible. This highlights how important it is to develop an automated and reliable system for seizure detection.

Electroencephalograms (EEGs), which record the electrical activity of the brain, are the primary tool used by neurophysiologists to identify seizures. EEGs are captured via electrodes placed on the scalp, measuring voltage differences between electrode pairs. This project, like a lot of research nowadays, aims to bridge the gap between clinical needs and technological advancements. In our case, we did this by automating the analysis of EEGs and attempting to predict whether an individual had a seizure or not. With the usage of automated systems for neonatal seizure detection we can increase the accuracy and efficiency of diagnoses. This in turn reduces the burden on health care professionals and ultimately saves lives and improves the quality of life for survivors.

High reliability was our main concern since we were developing a model capable of processing and predicting seizures. We leveraged modern machine learning techniques to our advantage and built upon previous results which addressed certain challenges regarding the classification of seizures. These challenges included but were not limited to, limited high-quality annotated data and variability in clinical recording settings. Our work contributes to the ongoing effort to make neonatal seizure detection more efficient and clinically viable.

We employed an extensive approach by combining time-series processing, supervised learning, and convolutional neural networks. We started by segmenting the EEG recordings into overlapping intervals to ensure sufficient training data. Through filtering, downsampling, and standardization we could remove noise from each segment and enhance the relevant signals for seizure detection. Having done this we utilized supervised learning techniques to train models using labeled datasets annotated by experts. We explored a few labeling strategies to evaluate the impact of annotation agreement on model performance. Having done this, we implemented CNNs as the core classifier where we took advantage of their ability to learn spatial and temporal patterns in EEG signals automatically. We evaluated our networks, after being trained, using key metrics such as AUC-ROC, precision, recall, and F1-score to ensure that they are effective and reliable.

Prior work

Seizure detection is not a new area of research. Extensive work has been previously done on the subject, especially in producing models that can perform accurate diagnoses promptly. Prior work by Davíð H. Ágústsson in his MSc thesis and Ana Borovac in her PhD thesis offered us valuable insight into the usage

of deep learning for this task. We utilized key findings from both works to contextualize our project and highlight significant advancements in the subject.

The relationship between seizure detection training methods and architectural design is examined in Ágústsson's thesis. He analyzed the EEG waves using deep neural networks. According to Ágústsson's research, techniques like ensembling, segment translations, and mixup produced positive outcomes, even though growing architectural complexity has limits when it comes to performance improvement. These techniques not only improved generalisation but also improved model prediction calibration, making them easier to understand in a clinical setting [1].

Borovac's dissertation presents an alternative method of approaching the issue. It tackles issues specific to this field, namely the lack of well-annotated datasets and the differences in recording conditions amongst NICUs. Her study emphasizes the value of using consensus annotations from numerous experts and how doing so increases the model's dependability. Additionally, Borovac shows how ensemble learning may be used to preserve model performance while yet complying with privacy laws that restrict data sharing. Moreover, Borovac investigated calibration techniques as a way to measure prediction uncertainty, which is a crucial aspect of clinical decision-making [4].

Collectively, these studies demonstrate how important it is to apply strong training techniques with domain-specific modifications when addressing such issues. Borovac's investigation into data quality and calibration is enhanced by Ágústson's work on novel training techniques. Their combined efforts offer a thorough framework for creating clinically relevant seizure detection systems.

Dataset and data processing

For this project, we used a dataset comprising 79 EEG recordings of newborn infants, collected at Helsinki University Hospital. Each recording lasted approximately 74 minutes and was sampled at a frequency of 256 Hz. The recordings included 21 channels, each corresponding to measurements from one of the 21 electrodes.

Additionally, an annotation dataset accompanied the recordings. In this dataset, each second of the corresponding EEG recording was labeled by three professionals: a value of 1 indicated that the second belonged to a seizure segment, while a value of 0 indicated no seizure activity.

The EEG data was processed using an EDF reader implemented with code adapted from the Visbrain project [5].

Methods

Data preprocessing

As each recording was processed each signal was filtered with a bandpass filter to remove noise. This was done using hard coded EEG filter using coefficients gotten from this repository. The signal was further downsampled from 256 Hz to 32 Hz to decrease the size of the input to our classifiers. Finally, each signal was standardized such that the mean was zero and standard deviation one. The standardization of a time series x(t) was performed as follows

$$z(t) = \frac{x(t) - \mu}{\sigma}$$

where μ is the mean of the signal and σ its standard deviation.

We decided to look at 16 second intervals of the EEG recordings with a 12s overlap as proposed by Ana Borovac in her thesis [4]. An episode was defined to be at least 10 seconds long and even though the average episode was much longer we chose not to lengthen our segments. A segment of length T seconds would mean that all seizure segments of length t seconds, where $10 \leq t < T$, would be discarded. The overlap of segments not only increased the number of training samples but also helped capture patterns in the data by allowing the model to observe the same events from slightly different perspectives.

As each segment had labels based on annotations by three experts we had to define how they would be labeled. We chose three different ways to label individual segments of EEG recordings, *consensus*, *majority* and *contains*. They were defined as

- *Consensus labeling*: The segment is labeled as true if each expert labels each second of the segment as true.
- *Majority labeling*: The segment is labeled as true if at least two of the three experts label each second of the segment as true.
- *Contains labeling*: The segment is labeled as true if at least one of the three experts labels each second of the segment as true.

For each labeling type a segment was then labeled as false if every expert labeled each second as false. Every other segment, that is each segment with conflicting labels, was discarded. Prior reasearch shows that keeping some of these conflicting segments can be beneficial for classification [4] but that possibility was not explored in our analysis.

For the training of the convolutional neural networks we further downsampled the majority class in the training set such that the amount of seizure segments and non-seizure segments were the same. This was done to address the major class disparity as well as ease to computations. When creating the training, testing and validation sets, the data was spilt on a patient basis, that is, if some segments came out of the same recording they would be in the same set.

Feature classifier

A feature classifier was chosen as a baseline model. For each segment we extracted features that captured the charectaristics of the time series. The following five amplitude features were chosen. First of all the three Hjorth parameters which are commonly used in analysis of EEG data [2]. Given a time series $y(t)$ the parameters are

$$\text{Activity} = \text{var}(y(t)), \quad \text{Mobility} = \sqrt{\frac{var\left(\frac{dy(t)}{dt}\right)}{var(y(t))}}, \quad \text{Complexity} = \frac{\text{Mobility}(\frac{dy(t)}{dt})}{\text{Mobility}(y(t))}$$

The parameters represent the variance, mean frequency and change in frequency and are therefore highly suitable for our applications. The Hjorth parameters were computed using the `pyeeg` package [3].

The other two features were the absolute band power in the [2-4] Hz frequency band and in the [4-6] Hz frequency band. EEG recordings are named based on their frequency range and the most commonly studied are delta (0.5 to 4Hz), theta (4 to 7Hz), alpha (8 to 12Hz), sigma (12 to 16Hz) and beta (13 to 30Hz) [9]. Research has shown that high amplitude delta oscillations corrolate strongly to generalized epileptic seizures [7], which motivated the selection of the [2-4] Hz range for the first frequency band. The [4-6] Hz range was then chosen to capture the transition from the delta range to the theta frequency range. The band power was implemented using code based on Raphael Vallats tutorial on bandpower [10].

Each of these features only applies to single channel signals unlike the multichannel signals in our dataset.

For each segment the amplitude features were therefore calculated for each channel. The mean of each feature over the channels was then taken resulting in five features for each segment. The mean of the features was chosen to account for seizures which show up on multiple channels.

An `XGBoost` binary classifier was trained on the extracted features data with 10-fold cross-validation. `XGBoost` uses gradient boosting and is, as an industry standard a good choice for our classification task. The model was implemented using `XGBclassifier` with default values except for scaling of the positive weights to account for the major disproportion in number of seizure and non-seizure segments.

Convolutional Neural Networks

The baseline convolutional neural network was simple and quite compact, loosely based on the baseline model presented by Davíð in his thesis [1]. The structure is similar but we did not flatten and gather channels or utilize channel-wise max pooling, instead training using only one ConvBlock and global average pooling. This model, along with all subsequent models, utilized early stopping activated when the validation loss plateaued. Binary cross-entropy loss and an output layer containing a single neuron with sigmoid activation were also consistently used in all models. This simple model achieved a surprisingly high AUC of 0.8701. However, there was still much work to be done. We also tried the same structure to train a 2D convolutional network but that turned out to be a really ineffective training method, achieving an AUC of only around 0.65. That is logical considering much research shows that one dimensional convolutions work better with time series data than 2 dimensional [6]. Consequently, all focus was put on constructing a more effective 1D convolutional neural network.

The next model contained 3 ConvBlocks, each of them with a different number of filters and kernel size.

- ConvBlock 1
  - Conv1D Layer: 64 filters, kernel size 16, ReLU activation, L2 regularization
  - Batch Normalization: Stabilizes and accelerates training
  - MaxPooling1D: Reduces dimensionality
  - Dropout: Rate of 0.5 to prevent overfitting
- ConvBlock 2
  - Conv1D Layer: 128 filters, kernel size 8, ReLU activation, L2 regularization
  - Batch Normalization
  - MaxPooling1D
  - Dropout: 0.5
- ConvBlock 3
  - Conv1D Layer: 256 filters, kernel size 4, ReLU activation, L2 regularization
  - Batch Normalization
  - Global Average Pooling

Furthermore we reduced the learning rate allowing for finer weight adjustments, which can be beneficial for a deeper network such as this one. We also added balanced class weights, giving higher weights to the minority class. We also increased our batch size from 32 to 64 in an attempt to improve training efficiency and stabilize gradient estimates. This model achieved an impressive AUC of 0.9299, stopping

due to validation loss plateau after 9 epochs. We were quite satisfied with this AUC but only achieved a recall of 0.85 which we were keen to get over 0.90.

Our final model is very similar to the one that came before with some slight changes to improve the recall. The first one was an alteration to the class weights, emphasizing the seizure class with a weight of 3 to 1 for non-seizures. This encourages the model to minimize false-negatives, therefore heightening recall. The most substantial change was the addition of threshold optimization based on our desired recall value of 0.90. This was done by computing the precision-recall curve, calculating the the difference between the recall values curve and our desired recall, and identifying thresholds that either meet or exceed our desired recall value of 0.90. This enabled us to reach our desired recall value without negatively affecting precision and F1-score too much, while also reaching an AUC of 0.9366.

Model evaluation

Evaluation of all classifiers was conducted using four key metrics: AUC-ROC, Recall, Precision and F1-score. These particular metrics provided a comprehensive view of model performance, especially helpful when working with an imbalanced dataset, giving us a good idea of the model's ability to distinguish between classes, its bias towards each class, and so on.

- *AUC-ROC (Area Under the ROC Curve)*: Represents the probability of a classifier ranking a randomly chosen seizure event higher than a non-seizure event. It is derived from ROC (Receiver Operating Characteristic) curve, plotting true positive rate against false positive rate at numerous thresholds. This demonstrates the model´s ability to distinguish between non-seizure and seizure events regardless of threshold values, integral information for evaluation of a binary classifier.
- *Recall*: Measures the proportion of correctly identified seizure events to total seizure events, defined as
$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$
effectively telling us the percentage of actual seizures correctly identified by the model. [8]
- *Precision*: Measures the proportion of correctly predicted seizure events to all seizure events, defined as
$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$
effectively telling us how large a percentage of positive seizure predictions turned out to be correct. [8]
- *F1-score*: Calculates a harmonic mean of precision and recall defined as
$$\text{F1-score} = \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$
providing information regarding the evaluation of trade-off between recall and precision, most often concerning threshold selection. [8]

Working with such a heavily imbalanced dataset using accuracy as an evaluation metric can be problematic. Seizures identified by all three experts account for only about 10.7% of the data, therefore simply guessing the negative class would result in a highly misleading accuracy of 89.3%. Based on that, we deemed accuracy as an unnecessary metric.
Due to the high stakes nature of our task of neonatal seizure detection a high value for recall was prioritized. An unidentified seizure can have a devastating effect on a newborn but the repercussions of the occasional false diagnosis are much less dramatic. Therefore the threshold was adjusted in order to achieve a high

value for recall. This must however be done within reason as too low of a threshold can result in the model identifying nearly every segment as a seizure. That causes an attractive recall value often close to 1.00 along with an enormous amount of false positives which is a highly undesirable outcome in real-world application. Finding a good balance is crucial, and these evaluation metrics provide us with all we need to achieve that.

Results

Feature classifier

A feature based dataset was extracted using the process described above for each type of labeling. The feature classifier was then trained and tested using 10-fold cross validation for each of the datasets.

10-fold cross validation for consensus labeling

	Precision (%)	Recall (%)	F1-score (%)
Non seizure	96.8	85.1	90.5
Seizure	27.0	65.6	36.1
AUC	0.8291		

Table 1. Classification report for consensus labeling

10-fold cross validation for majority labeling

	Precision (%)	Recall (%)	F1-score (%)
Non seizure	96.1	81.7	88.2
Seizure	28.4	68.0	39.2
AUC	0.8151		

Table 2. Classification report for majority labeling

10-fold cross validation for contains labeling

	Precision (%)	Recall (%)	F1-score (%)
Non seizure	92.7	77.6	84.3
Seizure	33.3	63.4	42.8
AUC	0.7616		

Table 3. Classification report for contains labeling

These results indicate that using consensus labeling is the best approach for further training of the neural networks. The majority approach gets marginally better recall and F1-score for detecting seizures but for now we prioritized AUC.

Convolutional Neural Networks

The raw EEG was split randomly into a training, validation and testing set on a patient basis using the consensus labeling. The results of the baseline convolutional neural network were following:

	Precision (%)	Recall (%)	F1-score (%)
Non seizure	96.0	93.0	95.0
Seizure	64.0	75.0	69.0
AUC	0.8701		

Table 4. Classification report for baseline CNN

The training finished after 8 epochs. The results were better than obtained from the feature classifiers but still not optimal. The recall for example, was 75% meaning that only 3 quarters of seizures are detected which would not be very useful in real world applications.

The final CNN model was run with the same random training, validation and testing split as above. Now focusing on maximizing the recall the results of testing the model were:

	Precision (%)	Recall (%)	F1-score (%)
Non seizure	98.3	86.2	91.9
Seizure	51.4	90.1	65.6
AUC	0.9366		

Table 5. Classification report for final CNN

The confusion matrix for this run was:

	Non-seizure	**Seizure**
Predict Non-seizure	6795	118
Predict seizure	1085	1147

Table 6. Confusion matrix for testing of final model

The training finished after 13 epochs. The final model achieved a much higher AUC of 0.9366 as well as a major increase in recall, that is from 75% to 90%. The model caught 1147 seizure segments while missing only 118.

Conclusion and future work

In this project we aimed to develop both a feature based classifier and a convolutional neural network to detect neonatal seizures in multichannel EEG recordings. Creating the feature based classifier proved a success as we achieved as high of an AUC as 0.829 1. This was accomplished without fine tuning hyper-parameters or optimizing the features so there might be substantial potential for improvement in the model. Putting time and resources into developing a more complex feature based classifier might though not be wise since the extremely simple baseline convolutional neural network showed much better results 4 in every metric that was observed.

The convolutional neural network was also a success. By much trial and error we managed to develop a model that showed signs of real world utiliy. By adding and adjusting layers in the CNN as well as prioratizing recall with class weights, an 0.9366 AUC and a 90.1% recall was accomplished. These results demonstrate the power of deep learning in handling complex EEG data and its potential for clinical applications.

Our results showed promising advancements in neonatal seizure detection, but it is far from a complete model. There are various things, that if we had sufficient resources and time, we could do in order to enhance our methods' applicability and robustness. There is quite the room for improvement but here are a few suggetions:

1. **Expanding the diversity of the dataset:**
   In our current model the dataset is quite limited. It is relatively small and homogeneous, so in the future we would look to expand the dataset and get EEG recordings from more diverse NICUs with different setups, and recording protocols.

2. **Exploring recurrent architectures:**
   We have already shown that CNNs perform strongly in capturing spatial and temporal patterns in EEG signal, but a completely alternative approach might yield more favorable results. One such method is long short-term memory networks (LSTMs). When modeling long-term dependencies, LSTMs might improve the model's capabilities to identify longer or more subtle seizure patterns. This is undoubtably one of the first things we would explore given the resources and time.

3. **Integrating into clinical practice:**
   If we allow ourselves to be extremely optimistic, the best way to enhance our model is to apply it in real-life situations. By collaborating wiht NICUs and implementing our model into EEG monitoring devices, we would both be able to prevent certain seizures, while also gathering critical information to further improve our model based on in-the-field testing.

4. **Attacking the main problem head on:**
   The main problem remains the same, false positives. If provided more resources and time, we would dedicate a substantial amount to tackling them head on. Keeping this in mind while also applying the other mentioned improvements would be the best course of action moving forward.

5. **Ethical and privacy considerations:**
   Being aware that we are working with sensitive data if of the utmost importance so addressing privacy concerns would be one of the tasks we would take on. We could do this using techniques such as differential privacy or federated learning, in order to make sure that we comply with all data protection regulations while also progressing with our model.

Collaboration

As a whole, the project was a collective effort, and as a team, we equally shared the workload. At its core, the main reason for the project's success is our extensive collaboration. Having said this, specific problems were divided among team members but through excellent communication and a strong emphasis on teamwork, we maintained a streamlined path throughout the entire process. Key aspects, such as data preprocessing were initially tackled by Jóhannes since he has extensive knowledge on the matter. At the same time, feature engineering was the first thing on the agenda for Stefán given that he has the most experience out of the group on the matter. Having said this, we approached everything as a team, and in no way was any of us burdened by disproportionate work. The additional work of individual team members was always integrated into the larger framework of the team's effort which reflected in our commitment to a shared ownership of this project. This approach to the project not only streamlined the workflow but also created an environment for mutual learning. To summarize all of this, we can say that our success in this project is due to our collective efforts and unified problem-solving.

Bibliography

[1] Davíð Hringur Ágústsson. "Deep Neural Networks for Seizure Detection: A Study on Training Strategies and Architectural Designs". MSc thesis. University of Iceland, 2023.

[2] Wissam H Alawee, Ali Basem, and Luttfi A Al-Haddad. *Advancing biomedical engineering: Leveraging Hjorth features for electroencephalography signal analysis.* Last accessed 17 November 2024. 2023. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10750318/.

[3] Forrest Sheng Bao, Xin Liu, and Christina Zhang. "PyEEG: An Open Source Python Module for EEG/MEG Feature Extraction". In: *Computational Intelligence and Neuroscience* Volume 2011 (2011).

[4] Ana Borovac. "Towards Clinically Useful Neonatal Seizure Detection Algorithms". PhD thesis. University of Iceland, 2024.

[5] Etienne Combrisson and contributors. *Visbrain: A Brain Visualization Suite in Python*. https://github.com/EtienneCmb/visbrain. Accessed: 2024-11-21. 2024.

[6] Álvaro Teixeira Escottá, Wesley Beccaro, and Miguel Arjona Ramírez. *Evaluation of 1D and 2D Deep Convolutional Neural Networks for Driving Event Recognition*. Last accessed 16 November 2024. 2022. URL: https://www.mdpi.com/1424-8220/22/11/4226.

[7] Joel Frohlich, Daniel Toker, and Martin M Monti. "Consciousness among delta waves: a paradox?" In: *Brain* 144 (8 2021), pp. 2257–2277.

[8] Steinn Guðmundsson. "Machine Learning". Lecture notes for REI505M, University of Iceland. Nov. 2024.

[9] Chetan S. Nayak and Arayamparambil C. Anilkumar. *EEG Normal Waveforms*. Last accessed 21 November 2024. 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK539805/#:~:text=However%2C%20the%20most%20frequently%20used,beta%20(13%20to%2030Hz)..

[10] Raphael Vallat. *Compute the average bandpower of an EEG signal*. Last accessed 20 November 2024. 2018. URL: https://raphaelvallat.com/bandpower.html.