Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

# Deep learning for detecting interictal EEG biomarkers to assist differential epilepsy diagnosis

## An explainable Artificial Intelligence approach to neurological signal analysis

Master's thesis

2020
Laura Gschwandtner, BA

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

# Deep learning for detecting interictal EEG biomarkers to assist differential epilepsy diagnosis

## An explainable Artificial Intelligence approach to neurological signal analysis

Master's thesis

| | |
|---|---|
| Study programme: | Cognitive Science |
| Field of study: | Computer Science |
| Department: | Department of Applied Informatics |
| Supervisor: | Prof. Ing. Igor Farkaš, Dr. |
| Consultant: | Franz Fürbass, PhD |

Bratislava, 2020
Laura Gschwandtner, BA


Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

# THESIS ASSIGNMENT

| | |
|---|---|
| **Name and Surname:** | Laura Pauline Gschwandtner |
| **Study programme:** | Cognitive Science (Single degree study, master II. deg., full time form) |
| **Field of Study:** | Computer Science |
| **Type of Thesis:** | Diploma Thesis |
| **Language of Thesis:** | English |
| **Secondary language:** | Slovak |

**Title:** Deep learning for detecting interictal EEG biomarkers to assist differential epilepsy diagnosis

**Annotation:** The diagnostic procedure of epilepsy is often tedious and requires long-term video EEG monitoring (VEM) to capture clear epileptiform discharges. Also, differentiating epileptic symptoms from e.g. psychogenic non-epileptic seizures is hardly possible without VEM. A few EEG biomarkers are known, which could be used in combination with deep learning methods to enhance faster diagnosis of epilepsy from short interictal EEG recordings.

**Aim:** 1. Develop a deep neural network model for classification of interictal EEG recordings into epileptic and non-epileptic subjects.
2. Use machine learning to search for subtle biomarkers in routine-EEG, that could indicate chronic epilepsy.

**Literature:**
Engel J., Bragin A., & Staba R. (2018). Nonictal EEG biomarkers for diagnosis and treatment. Epilepsia Open, 3(Suppl 2), 120-126.
Fu X. et al. (2018). Negative effects of interictal spikes on theta rhythm in human temporal lobe epilepsy. Epilepsy & Behavior, 87, 207-212.
Van Leeuwen K. et al. (2019). Detecting abnormal electroencephalograms using deep convolutional networks. Clinical Neurophysiology, 130(1), 77-84.
Bagheri E. et al. (2019). A fast machine learning approach to facilitate the detection of interictal epileptiform discharges in the scalp electroencephalogram. Journal of Neuroscience Methods, 326, 108362.

| | |
|---|---|
| **Supervisor:** | prof. Ing. Igor Farkaš, Dr. |
| **Consultant:** | Franz Fürbass, PhD. |
| **Department:** | FMFI.KAI - Department of Applied Informatics |
| **Head of department:** | prof. Ing. Igor Farkaš, Dr. |
| **Assigned:** | 21.01.2020 |
| **Approved:** | 31.01.2020 |

prof. Ing. Igor Farkaš, Dr.
Guarantor of Study Programme

.....
Student

.....
Supervisor


Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

# ZADANIE ZÁVEREČNEJ PRÁCE

**Meno a priezvisko študenta:** Laura Pauline Gschwandtner
**Študijný program:** kognitívna veda (Jednoodborové štúdium, magisterský II. st., denná forma)
**Študijný odbor:** informatika
**Typ záverečnej práce:** diplomová
**Jazyk záverečnej práce:** anglický
**Sekundárny jazyk:** slovenský

**Názov:** Deep learning for detecting interictal EEG biomarkers to assist differential epilepsy diagnosis
*Hlboké učenie na detekciu interiktálnych biomarkerov EEG na pomoc diferenciálnej diagnostike epilepsie*

**Anotácia:** Diagnostická procedúra na epilepsiu je často zdĺhavá a vyžaduje si dlhodobé vizuálne monitorovanie EEG signálu (VEM), aby sa dali zachytiť jasné epileptiformné výboje. Taktiež, odlíšenie epileptických symptómov od napr. psychogenických neepileptických záchvatov je takmer nemožné bez použitia VEM. Je známych niekoľko EEG biomarkerov, ktoré by sa dali použiť v kombinácii s metódami hlbokého učenia s cieľom vylepšiť rýchlu diagnostiku epilepsie z krátkych interiktálnych EEG záznamov.

**Cieľ:** 1. Nadizajnujte model hlbokej neurónovej siete na klasifikáciu interiktálnych EEG záznamov na epileptické a neepileptické subjekty.
2. Pomocou strojového učenia nájdite jemné biomarkery v rutinnom EEG signáli, ktoré by mohli byť indikátorom chronickej epilepsie.

**Literatúra:**
Engel J., Bragin A., & Staba R. (2018). Nonictal EEG biomarkers for diagnosis and treatment. Epilepsia Open, 3(Suppl 2), 120-126.
Fu X. et al. (2018). Negative effects of interictal spikes on theta rhythm in human temporal lobe epilepsy. Epilepsy & Behavior, 87, 207-212.
Van Leeuwen K. et al. (2019). Detecting abnormal electroencephalograms using deep convolutional networks. Clinical Neurophysiology, 130(1), 77-84.
Bagheri E. et al. (2019). A fast machine learning approach to facilitate the detection of interictal epileptiform discharges in the scalp electroencephalogram. Journal of Neuroscience Methods, 326, 108362.

**Vedúci:** prof. Ing. Igor Farkaš, Dr.
**Konzultant:** Franz Fürbass, PhD.
**Katedra:** FMFI.KAI - Katedra aplikovanej informatiky
**Vedúci katedry:** prof. Ing. Igor Farkaš, Dr.

**Dátum zadania:** 21.01.2020

**Dátum schválenia:** 31.01.2020

prof. Ing. Igor Farkaš, Dr.
garant študijného programu

**DECLARATION**

I hereby declare that I elaborated this diploma thesis independently. Formulations and ideas taken from other sources are cited as such.

| | |
|---|---|
| Laura Pauline Gschwandtner | Bratislava, 07.08.2020 |

**Acknowledgements:** I want to thank the whole research group of Biosignal Processing at the AIT Austrian Institute of Technology for including me. I especially want to thank Manfred Hartmann and Franz Fürbass for their guidance over the past year and for teaching me everything I needed to know to make this thesis possible. I am very grateful for two insightful and exciting years of being part of the MEi:CogSci family in Bratislava and Vienna and want to thank Igor Farkaš for his continuous support. Further, I want to thank my parents for teaching me curiosity and Joey for providing me with the essential coffee and pizza during the writing process.

# Abstract

This project explores novel approaches to assessing electrical brain activity recorded through electroencephalography (EEG) for epilepsy diagnosis. The thesis at hand presents a machine learning algorithm for automatic classification of EEG recordings into epileptic and non-epileptic. In epilepsy patients, relatively short EEG recording periods without apparent epileptic events like seizures usually still show patterns different from non-epileptic EEG activity. Such interictal epileptiform discharges (IEDs) can serve as EEG biomarkers. However, to date no reliable and unambiguous diagnostic analysis solely based on the known IEDs has been established. Identifying additional interictal biomarkers for diagnosis could contribute to faster procedures and help to prevent misdiagnoses and ineffective or harmful treatment. Faster diagnostic procedures could furthermore save essential resources and make hospital beds available in urgent situations.

The presented thesis tackles this issue by making an end-to-end machine learning algorithm search for distinct patterns in EEG data and learn to use them to differentiate epileptic from non-epileptic recordings. By exploring means of explainable AI (XAI), the goal was to make the detected patterns explicit. Creating such explanations for algorithmic decisions further carries high ethical relevance in the health care field. Practitioners will only gain trust in a system and agree to use it, if the automated decisions are logical and the reasoning process can be explained clearly. This level of interpretability is not inherent in the deep neural network algorithms applied. Thus, within the scope of this work the concept of explainable AI was explored and implementation was attempted.

The presented study shows that automatic EEG analysis using deep learning is feasible. At the same time, extensive further research will be necessary to create clinically applicable, highly accurate and transparent algorithms for epilepsy diagnosis from routine EEG.

**Keywords:**

EEG, Epilepsy, Artificial Intelligence, Neural Networks, CNN, ResNet, Explainability

# Abstrakt

Tento projekt skúma nové prístupy k hodnoteniu elektrickej mozgovej aktivity zaznamenanej pomocou elektroencefalografu (EEG) na diagnostiku epilepsie. Táto práca predstavuje algoritmus strojového učenia pre automatickú klasifikáciu záznamov EEG na epileptické prípady a neepileptické. U pacientov s epilepsiou relatívne krátke úseky záznamu EEG bez zjavných epileptických udalostí, ako sú záchvaty, stále vykazujú vzorce odlišné od epileptickej aktivity EEG. Takéto interiktálne epileptiformné výboje (IED) môžu jednotlivo slúžiť ako EEG biomarkery. Doteraz však nebola stanovená žiadna spoľahlivá a jednoznačná diagnostická analýza založená iba na známych IED. Identifikácia ďalších interiktálnych biomarkerov na diagnostiku by mohla prispieť k rýchlejším postupom a pomôcť predchádzaniu nesprávnym diagnózam a neúčinnej alebo škodlivej liečbe. Rýchlejšie diagnostické postupy by navyše mohli ušetriť základné zdroje a sprístupniť nemocničné lôžka v naliehavých situáciách. Predkladaná práca sa venuje tejto problematike tak, že algoritmus strojového učenia typu end-to-end hľadá odlišné vzory v dátach EEG a naučí sa ich používať na rozlíšenie epileptických a neepileptických záznamov. Cieľom prieskumu pomocou vysvetliteľnej umelej inteligencie (XAI) bolo objasniť zistené vzorce. Vytváranie takýchto vysvetlení pre algoritmické rozhodnutia má naďalej vysoký etický význam v oblasti zdravotnej starostlivosti. Praktizujúci lekári získajú dôveru v systém a budú súhlasiť s jeho používaním, iba ak sú automatizované rozhodnutia logické a proces odôvodnenia môže byť jasne vysvetlený. Vytváranie takýchto vysvetlení nie je prirodzene možné pre algoritmy hlbokej neurónovej siete. V rámci tejto práce sme preto skúmali koncepciu XAI a pokúsili sa o implementáciu. Predložená štúdia ukazuje, že je možná automatická analýza EEG pomocou hlbokého učenia. Súčasne bude potrebný ďalší rozsiahly výskum na vytvorenie klinicky použiteľných, vysoko presných a transparentných algoritmov na diagnostiku epilepsie z rutinného EEG.

**Kľúčové slová:**

EEG, epilepsia, umelá inteligencia, neurónové siete, CNN, ResNet, vysvetliteľnosť

# Contents

# List of Figures



# List of Tables

# Abbreviations

**AED** Antiepileptic Drug

**AI** Artificial Intelligence

**ANN** Artificial Neural Network

**AP** Attention Pooling

**CNN** Convolutional Neural Network

**EEG** Electroencephalography

**HFO** High Frequency Oscillation

**IED** Interictal Epileptiform Discharge

**MDR** Multi-Drug-Resistant

**ML** Machine Learning

**PNES** Psychogenic Non-Epileptic Seizures

**ResNet** Residual Network

**VEM** Video-EEG Monitoring

**XAI** Explainable Artificial Intelligence

# 1 Introduction

This thesis explores how machine learning algorithms can be applied to assess electroencephalographic signals (EEG) for improving epilepsy diagnosis. The current state of the art in differential epilepsy diagnosis is a tedious and time-consuming process (Baumgartner & Pirker, 2019). Additionally to a faster identification of potential epilepsy patients, an improved diagnostic procedure could rule out non-epileptics from lengthy subsequent examination and prevent false, ineffective treatment with physical and psychological side effects (Doss & LaFrance, 2016).
The research questions tackled by the study at hand are built up on each other in the following manner:

- Can artificial intelligence (AI) be applied to reliably classify interictal EEG recordings into epileptic and non-epileptic subjects?
- Can machine learning (ML) identify epilepsy biomarkers in interictal EEG, which are not/ hardly visible through human inspection?
- Can this AI algorithm be made explainable and transparent to ethically assist medical diagnosis?

To attempt answering the first question, several machine learning algorithms are created and compared. Hereby variations of the sub-type of convolutional neural networks (CNNs) are employed. Those supervised learning algorithms can be very efficient and powerful when dealing with large amounts of complex data. The key objective of this thesis is to train such classifier networks to automatically differentiate between epileptic and non-epileptic EEG recordings. Contrary to previous machine learning approaches which mainly focused on quicker detection of clearly visible epileptic EEG patterns or surgical management (Abbasi & Goldenholz, 2019), the approach here is to create an end-to-end system to directly perform a binary classification on raw interictal routine EEG recordings. An interictal period defines the time between two epileptic seizures, where the patient usually is not experiencing any direct symptoms of the disease. However, routine EEG recordings during such time periods were still found to exhibit different patterns than non-epileptic EEG (Engel et al., 2018). If such interictal periods could serve as solid diagnostic measurements, the current diagnostic procedure of epilepsy could potentially be immensely accelerated.

The second research question is again motivated by the fact that interictal EEG signals of epilepsy patients can show subtle differences (interictal epileptiform discharges (IEDs)) compared to non-epileptic EEG activity which could potentially serve as diagnostic biomarkers. Most IEDs that are known to date and quite clearly visible are not

completely reliable for diagnosis. They can also be present in non-epileptic EEG and do not necessarily occur in epileptics. Hence, the study at hand does not put too much focus on specific, known IEDs. Instead, this project aims for an unbiased learning process to enable the machine learning algorithm to base its decisions on any distinctive EEG patterns it detects. Explainability techniques are then used to analyze the algorithms inner workings and make its decision making processes apparent and visible. These tools can be used to point to patterns in the data the algorithm bases its classification decision on. The hope is, that this can offer new insights and identify novel interictal EEG biomarkers. Integrating additional interictal biomarkers could enhance the diagnostic procedure and make it less prone to misdiagnoses.

The evaluation of different explainability tools also targets the third research question and aims to make the proposed algorithm transparent and understandable. This thesis suggests that algorithms for medical and in particular diagnostic purposes should generally be required to be explainable. Hence, this study treats explainability as a core value when creating clinical software and therefore attempts to implement it. It can be assumed that medical practitioners will only gain trust in a system and agree to use it, if the automated decisions are logical and ideally the reasoning process can be explained clearly. Unfortunately, this is not inherently the case in neural network algorithms[1]. As such networks are very powerful and able to accomplish specific tasks much better than traditional, more transparent approaches, they will continue to be used extensively, also in the ethically complex healthcare field. Therefore, within the scope of this work the concept of explainable AI (XAI) will be explored and the implementation of concrete methods to improve the proposed algorithm will be attempted and evaluated.

## 1.1 Interdisciplinarity

The proposed topic is highly interdisciplinary, involving perspectives ranging from neurological signal processing to medical ethics. The relevant disciplines included can be broken down into the main basic fields neuroscience, psychology, artificial intelligence and philosophy.

The neuroscience perspective is important for in-depth understanding of epilepsy as a neurological disease and how it affects the brain and its neurons and connections. This comprehension is necessary to be able to build a good classification algorithm, considering the state of the art in diagnosis, treatment and neurological understanding of the disease and potential challenges and goals. Understanding the correlations between epilepsy and mental health and the psychological burden of the disease and its treatment is an integral part of working with this topic. Further, some psychiatric disorders can show ambiguous

[1]Details on functionality of those networks and the advantages and drawbacks will be discussed in further chapters of this thesis.

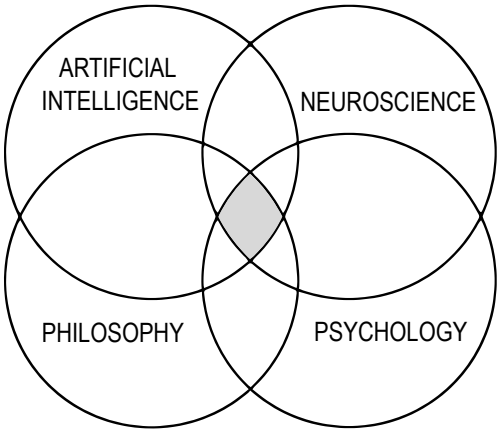


Figure 1: Interdisciplinary positioning of the proposed research study.

symptoms similar to epileptic seizures and their differentiation is essential for reliable diagnosis. Artificial intelligence methods are utilized in this thesis project for the main goal of gaining more insight on how to facilitate and speed up epilepsy diagnosis. AI hereby helps to search for patterns in large amounts of EEG data, that are opaque and hardly visible for humans. The implemented technology therefore constitutes the major focus of this thesis. The field of philosophy comes into play when ethical implications of applying AI to medicine are discussed. Creating autonomous technology for healthcare responsibly and in an unbiased manner should constitute a key value in this fast evolving area. This results in the objective of creating AI systems that are somewhat explainable and transparent to the developer and also to the end-user (e.g. medical professionals). Thus, the ethical part of this thesis is closely linked to the methodological implementation and the AI technology in use.

## 1.2 Positioning within cognitive science

Within the wide scope of (the) cognitive science(s), approaching a research question through the use of artificial neural networks (ANNs) can traditionally be theoretically embedded within the connectionist paradigm. However, this section quickly walks through the historical interaction of cognitive science and AI with a focus on explainability and relates it to the research topic at hand and the research questions posed.

Historically, the breakthrough of the theoretical interdisciplinary overlap of cognitive science and AI can be dated back to Hilary Putnam and Jerry Fodor's Computational Theory of Mind (CTM) (Fodor, 1975). CTM coined the belief that the mind works in very much similar ways like a digital computer and this theory suggests a crossover effect. According to this theory, the failure of creating efficient and functional AI from theorized models of the mind prompts also the model of the mind to be implausible. On the other

hand, being able to create AI systems capable of human cognitive capacities suggests that those would also constitute a plausible theory of the mind (Westberg et al., 2019). In the end, what computationalism can contribute to this thesis is that using artificial neural networks inspired by the human brain's neurons promises to combine the human capability of learning concepts and patterns from data with the computer's processing power and its ability to detect details that are invisible for humans.

In the 1990s the interest in computationalism declined and a new paradigm called connectionism gained popularity. Connectionism tries to use artificially built neural networks to explain the workings of the human mind and brain. On the other hand, computer scientists were hoping to create better, faster and more flexible systems by creating technologies that are strongly inspired by the human brain and its neuronal connections. Convolutional neural networks for example were inspired by visual cortical neurons of mammals and residual networks (ResNets) got their basic concept from pyramidal cells in the cerebral cortex. Using convolutional layers as well as ResNet structure, the network utilized in this thesis project integrates two different neuron types. However, the diagnostic purpose of this research is different from the basic connectionist goals of gaining deeper understanding of the mind and brain. This goal could be brought in line with connectionism by approaching this topic through computational simulation of the epileptic brain using neural networks. However, this will not further be explored in this thesis.

The next major cognitive paradigm emerged through the embodied cognition movement. Admittedly this paradigm does not offer much to AI applied in a context like presented here and even contradicts it in a way.

The integration of ethical considerations and explainability into this thesis broadens its standing within cognitive science. As can be deduced from Westberg et al. (2019), cognitive science is still unable to offer basic practical advice, even though cognitive scientists show growing interest in the field of XAI. Unfortunately it seems that even though the research fields of AI and cognitive science are strongly intertwined and share historical context they do not speak the same language. Cognitive science is concerned with the high-level challenges of AI, wondering if general-level artificial intelligence will one day emerge and debating how to deal with singularity. In the meantime AI researchers on the practical side are trying to implement explainable machine learning methods on a very fundamental level. Here, the question is not (yet) what neural and psychological mechanisms underlie understanding or how to generate embodied empathetic explanations. The urging question currently is how any kind of relevant information on the sense making process can be extracted from inside the hidden layers of a deep artificial neural network. To achieve this, insights and know-how from cognitive science, neuroscience and psychology could potentially improve the comprehension of the inner workings of artificial neural networks. Knowledge about how explanation and understanding of complex decision processes work in the human brain could provide new ideas on how to improve ANNs.

Collaboration between disciplines can be a very profitable approach, but the premise for a successful exchange of ideas is to meet at the same level of debate and to first gain understanding of each others knowledge and progress.

On a final note of this subsection I want to position the approach taken in this thesis within the theoretical beliefs of predictive processing and connectionism. An enactive approach seems too far-fetched for a narrow AI system as proposed here. It mainly builds up on the connectionist idea of neural networks including the concepts of backpropagation and predicting the given environment like suggested in predictive coding.

## 1.3 Thesis outline

Chapter 2 lays the foundation of this thesis by explaining all components needed for thorough understanding of the problem statement. The essential theoretical background can thus be split into three basic elements. In Section 2.1 a definition of epilepsy and its symptoms is given and the current state of the art in epilepsy diagnosis is described. As a second component Section 2.2 explains the functionality of electroencephalography (EEG) for recording and analysing brain activity. Finally, Section 2.3 discusses artificial intelligence as a tool for handling large amounts of (medical) data and how such tools can be handled in a transparent and trustworthy manner.

In Chapter 3 the methods applied in the study at hand are explored in detail. First, Section 3.1 describes the data set used for training the algorithms. In Section 3.2 the different components utilized for building the algorithm and creating explainability are listed and explained. Section 3.3 goes into various experiments that were conducted using the different methodological components. Finally, Section 3.4 explains the metrics used to evaluate the performance of the algorithms. Hereby terms such as accuracy, sensitivity, specificity and F1 score are introduced.

The final results are presented in Chapter 4. First, an overview of the results achieved with different model setups is given. Performance variances due to changes in various parameters like input data, network type etc. are explored. In the end the results of exploring XAI methods are layed out in Section 4.1.

The achieved results and their implications are discussed in Chapter 5. The impact that different experimental changes had on the outcome are reviewed. Limitations and potential implications of the study for the field of cognitive science are debated.

Finally, Chapter 6 points to the main contents and insight of this thesis and poses some concluding remarks. Further, discussed limitations and ideas for improvements and future research are summed up.

# 2 Theoretical background

## 2.1 Epilepsy

Epilepsy is amongst the most common neurological diseases currently affecting approximately 50 million people worldwide. While in high-income countries 49 out of 100 000 people are estimated to be diagnosed with epilepsy per year, in low- and middle-income countries up to 139 per 100 000 are diagnosed. It is estimated that almost 80% of epilepsy patients live in low- and middle-income countries. This imbalance is likely due to a higher prevalence of infectious diseases like neurocysticercosis or malaria, increased birth complications or road accidents leading to head injury and brain damage and a lack of accessible healthcare. Additional potential causing factors of epilepsy include genetic conditions, stroke and brain tumors (World Health Organization, 2019).

The condition is characterized by recurrent seizures created through spontaneous synchronous electric discharge of neurons in the patient's brain. However, clearly characterizing this disease has been tough and many definitions were proposed throughout the last two decades. The International League Against Epilepsy (ILAE) 2005 defined epilepsy as *"a disorder of the brain characterized by an enduring predisposition to generate epileptic seizures and by the neurobiologic, cognitive, psychological, and social consequences of this condition. The definition of epilepsy requires the occurrence of at least one epileptic seizure"* (Fisher et al., 2005). In 2014, ILAE added an *operational (practical) clinical definition of epilepsy* in more detail to meet the clinical needs of medical practitioners working with epilepsy diagnosis and treatment (Fisher et al., 2014). Hereby, epilepsy is defined as follows:

> "Epilepsy is a disease of the brain defined by any of the following conditions
>
> 1. At least two unprovoked (or reflex) seizures occurring >24 h apart
> 2. One unprovoked (or reflex) seizure and a probability of further seizures similar to the general recurrence risk (at least 60%) after two unprovoked seizures, occurring over the next 10 years
> 3. Diagnosis of an epilepsy syndrome
>
> Epilepsy is considered to be resolved for individuals who had an age-dependent epilepsy syndrome but are now past the applicable age or those who have remained seizure-free for the last 10 years, with no seizure medicines for the last 5 years." (Fisher et al., 2014)

Seizures, also referred to as ictal events, can lead to one or more symptoms like the impairment or loss of consciousness, clonic or tonic muscle movement or staring. Epileptic

seizures can be classified based on their onset location into the three gross types focal, generalized and unknown, which again each can be subdivided into motor and non-motor seizures (Fisher et al., 2017). In 2017 the ILAE further porposed a general framework for the classification of epilepsies, considering seizure and epilepsy types, co-morbidities and causes, as visible in Figure 2 (Scheffer et al., 2017).

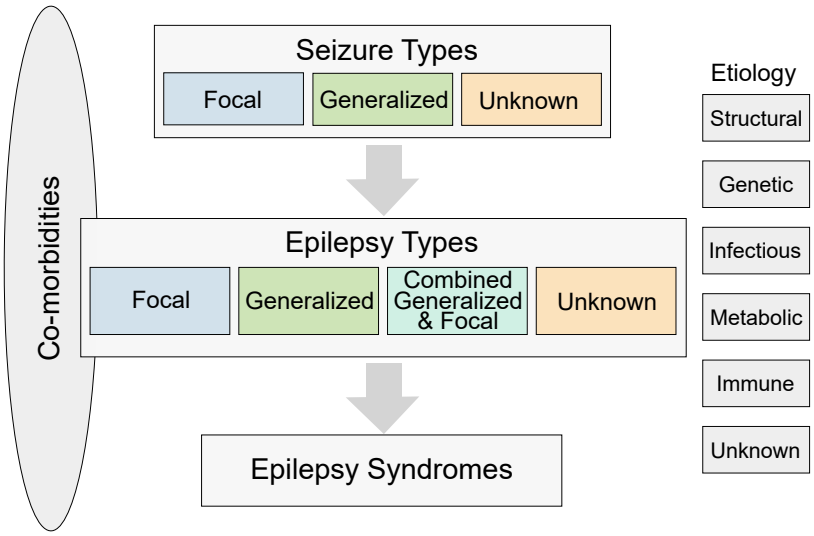


Figure 2: General framework for epilepsy classification, drawn after image from Scheffer et al. (2017).

The presented definitions display that epilepsy is a highly diverse disease, which can manifest itself in various ways. Anyhow, seizures often are a major cause for the patient's suffering. Depending on the place and time of occurrence and the type of seizure, those events can be very dangerous and cause serious injuries. Additionally, people suffering from epilepsy have to deal with stigma and discrimination in many regions across the world. This often leads to reduced opportunities in education, personal freedom, insurance or even prohibition of occupation or marriage. The cumulative incidence of mental problems in drug-resistant partial epilepsy patients referred to epilepsy surgery centers is notably high, ranging from 50% to 80% (Luders, 2008).

### 2.1.1 Diagnosis

The current common diagnostic procedures consist of different measures and activities. When epilepsy is suspected, medical professionals first perform a detailed anamnesis to find out about the medical history, symptoms and social context of the patient. Ideally, this acquired information already suggests the most likely epilepsy syndrome. As epilepsy is defined by abnormal neuronal activity, it is common practice to additionally record electrophysiological activity of the patient's brain through electroencephalography (EEG).

This is first done via routine EEG recordings of approximately 30 minutes. The use of EEG is often essential to assert a first clear diagnosis. If those measures are not effective and a clear diagnosis cannot be made, mobile long-term EEG or long-term video-EEG monitoring (VEM) is performed. Hereby the primary goal is to record ictal activities (seizures) and diagnose the patient according to those observations. This is the most direct and reliable means to diagnose epilepsy. However, long-term monitoring is preferably avoided if possible due to the lengthy and tedious procedure. A schematic overview of the use of EEG in the diagnostic procedure is depicted in Figure 3.

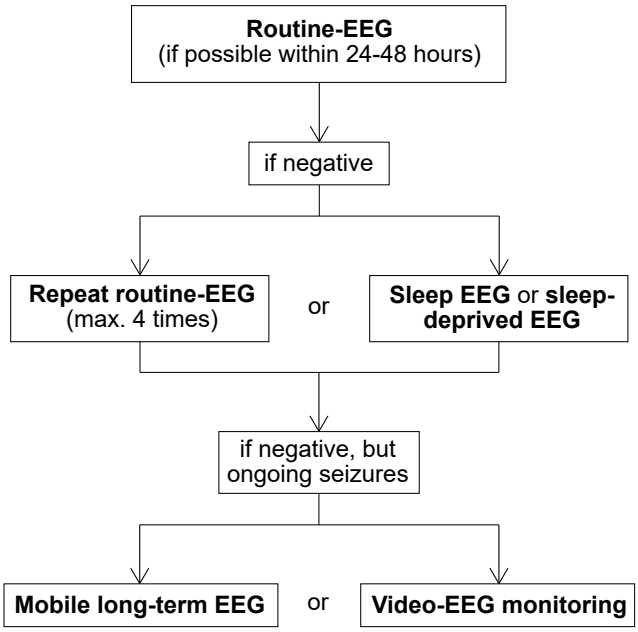


Figure 3: Schematic overview of EEG use in epilepsy diagnosis, adapted and translated from Baumgartner (2001).

Furthermore, neuroimaging techniques like Magnetic Resonance Imaging (MRI) or computed tomography (CT) scans are usually utilized to identify structural abnormalities.

Even though the occurrence of epileptiform seizures is the main characteristic of this disease, the diagnosis cannot rely solely on the occurrence of such ictal events. Seizures usually do not follow strict patterns. Patients can be seizure-free for weeks or even months and still generate new ictal events at times. On the other hand only about 30% of people, who have had a single ictal event will develop further seizures and thus indicate chronic epilepsy. Therefore, anticipating if and when a next seizure might happen is hardly possible and not generalizable. Nevertheless, experts often conclude that the ability to predict the development of ictal events (ictogenesis) would provide the most valuable advancement to create effective epilepsy treatment (Engel et al., 2018).

Besides seizures, interictal epileptiform discharges visible in EEG recordings also serve as biomarkers for epilepsy diagnosis. Those IEDs can be spotted by medical professionals

or specialized software. Unfortunately, such discharges are not deterministic and do not occur reliably in all epileptics alike. Thus, neither the observance of a single ictal nor of interictal events in the EEG can assure correct diagnosis of chronic epilepsy and certainly is not a clear predictor of ictogenesis.

Due to the various ambiguous factors stated above, epilepsy is often misdiagnosed. Subsequently, misdiagnosed patients are either treated for an illness they do not have (false positive) or people who are ill are not treated (false negative). Naturally, both of those diagnostic error types can be very harmful to the people affected and exceedingly expensive for the healthcare system. Accordingly, finding diagnostically more conclusive biomarkers is still a big challenge in creating more reliable diagnostic procedures and on-point treatment. Research on EEG interpretation reliability highlights, that the task of identifying abnormal EEG recordings shows strikingly low inter-rater agreement of 55% among qualified professionals (Grant et al., 2014). As discussed, those difficulties in reliably evaluating EEG data for epilepsy diagnosis have several adverse effects for patients and healthcare systems. Additionally, the lengthy diagnostic procedure complicates epilepsy research studies, which again decelerates the process of testing and bringing out new epilepsy medication or developing other scientifically verified treatment possibilities.

Depending on the specific use case, the demands on an automatic diagnosis system strongly vary. Generally speaking, a diagnostic measure which tends to mistakenly identify a lot of false positives comes with a high cost. This tendency can lead to physician not trusting the system and ignoring results. Given the present study, data set and research question, false positives result in potential inappropriate treatment with anti-epileptic medication and other complications. Undoubtedly though, false negatives are also highly unwanted here. This would result in sending an epileptic person home from the hospital, because he or she has been mistakenly claimed healthy by the diagnostic system. Furthermore, this means epileptics classified healthy could be mistakenly included in drug studies or unknowingly face situations that can be dangerous for them or induce more seizures.

#### 2.1.1.1 Monitoring and provocation

Long-term monitoring of potential epilepsy patients constitutes an essential part of state-of-the-art epilepsy diagnosis. Inpatient long-term video-EEG monitoring (VEM) for a final diagnosis is usually performed at dedicated epilepsy monitoring units (EMUs), while mobile long-term EEG can be conducted at the patient's home. Those procedures are important to record clear epileptic activities, specifically ictal events (seizures), which are hardly ever present in routine EEG (Baumgartner, 2001). In the case of VEM, simultaneous behavioural symptoms during seizures are captured on video. Further, interictal epileptiform discharges (IEDs = spikes, spike bursts, sharp waves) are also captured with

higher probability in long-term monitoring. Thus, these procedures offer evidence for quite reliably differentiating between epileptic and non-epileptic seizures. Further it can enable the classification of seizure types. For patients who do not respond to any medication, surgery often remains the only hope. In the case of such treatment-resistant epileptics, monitoring eventually becomes essential for pre-surgical evaluation. In case no seizures occur during a few days of VEM, they often have to be purposely provoked. Such provocation procedures commonly include sleep deprivation and reduction of regular anti-epileptic drugs (AEDs). If those traditional measures are ineffective, exercise or specific stimuli can be applied to provoke epileptic events (Baumgartner & Pirker, 2019).

#### 2.1.1.2 Differential diagnosis

The main challenge in differential diagnosis of epilepsy is to distinguish it from Psychogenic non-epileptic seizures (PNES) and Organic non-epileptic seizures (NES). Both diseases show symptoms that appear very similar to epileptic seizures, but have another underlying cause. This different origin requires completely different treatment and misdiagnosis can be harmful.

At least 10-40% of potential epilepsy patients are diagnosed with psychogenic non-epileptic seizures (PNES) after long term monitoring and many others instead are misdiagnosed with epilepsy and treated with futile, but potentially harmful medication (Doss & LaFrance, 2016). PNES show similar symptoms as epileptic seizures with some significant differences, that can vary and are sometimes hard to identify clearly. PNES can be a symptom of different psychiatric or psychological issues. It is assumed that stress and several other factors interfere and escalate into provoking a seizure-like event. Crucial determinants of developing PNES seem to be adverse childhood events, traumas, problematic parent-child relationships and other early influences on later stress behaviour and personality development. Further, unfortunate or stressful current living conditions have a favourable effect on the development of the disease. Many patients additionally suffer from other psychiatric illnesses, like PTSD (post-traumatic-stress disorder) or personality disorders like borderline disorder. The only known effective treatment for PNES is psychotherapy and dealing with potential additional psychiatric issues. Symptomatically PNES can be distinguished from epilepsy due to patients having closed instead of open eyes during a seizure, a longer duration ($>$ 5min) than commonly seen in epileptic seizures or irregular motor activity. Nevertheless, the main and most obvious indicator is, that PNES occur without any epileptiform electrical activity visible through EEG recordings during those seizure-like events. Usually also significantly less or no spikes or other epilepsy biomarker occur in PNES patients. Therefore, the most reliable diagnostic measure currently is to apply long-term VEM and wait for a seizure to be captured (Devinsky et al., 2011). As this is rarely done right away for every potential epilepsy patient

due to limited time and resources, there is a considerably high risk for patients suffering from PNES to be misdiagnosed with epilepsy and treated falsely before the mistake is recognized.

Part of the literature on differential epilepsy diagnosis also mentions organic non-epileptic seizures (NES). Just like PNES, organic NES constitute seizures which are not caused by rhythmic neuronal discharges in the brain typical to epilepsy. Unlike psychogenic seizures, organic NES are induced by physiological conditions that can be both neurological or non-neurological. Neurological causes can range from sleep disorders and cerebrovascular disorders to movement disorders. Non-neurological causes include metabolic abnormalities, cardiac arrhythmia or toxic ingestion (Hopp, 2019), potentially sometimes even elicited through anticonvulsant drug toxicity (Weaver, 2004).

Especially for distinguishing PNES and NES from epilepsy it could be a big improvement to develop a faster alternative of merely evaluating a rather short EEG recording pieces, like aimed at with this thesis.

### 2.1.2 Prevention and treatment

Epileptogenesis could be prevented in approximately 25% of cases beforehand. Hereby, the most important action is to avert head injury and brain damage. Thus appropriate perinatal care as well as reducing car crashes should be priorities. Further, cardiovascular risk factors as well as infections of the central nervous system should be reduced in order to prevent the development of new cases of epilepsy (World Health Organization, 2019). The treatment of epileptic patients mainly aims to control and prevent the development of further seizures.

#### 2.1.2.1 Medication

According to the WHO, given the right antiseizure medication up to 70% of epileptics could become seizure free (World Health Organization, 2019). Finding a drug and dosage that fits the patient can be a tough task though. There are several different antiepileptic drugs (AEDs) currently on the market, which have all been found effective for different people, depending on the type of epilepsy, interference with other medication in use and probably some other partly unknown factors. Table 1 provides an overview of the most common anti-epileptic drugs and their impact on patients with specific seizure types (Knezevic & Marzinke, 2018). Broadly speaking there are three possible outcomes during the anticonvulsant drug titration. Ideally, the patient can become seizure free with no severe side effects. The second possibility is that the patient will show serious adverse reactions to the AED in use and will not be able to continue medication. Third, even though the patient is compliant and the dosage has already been increased to the recommended maximum, the patient will not show any reduction in seizure occurrence. In

the second and third case, different antiseizure medication can be tested separately or added to the previous AED. If none of these measures show effect within a six months to two year period of unsuccessful trials (depending on the severity of the illness), the patient can usually be diagnosed with pharmacoresistant epilepsy. Beside patients never responding to drug treatment from the start, this status, also called multi-drug-resistant epilepsy (MDR), can even unsuspectedly be reached after years of successful medication. While the reasons for developing MDR epilepsy are unclear, oftentimes surgery remains the only chance for the patient to become seizure free (Luders, 2008).

Table 1: Overview of seizure types and suitable medication, adapted from Knezevic & Marzinke (2018).

| **Seizure type** | **Presentation** | **First-line drug options** | **Adjuvant or alternative drug options** | **Drugs that may worsen seizure** |
|---|---|---|---|---|
| Generalized tonic clonic | Initial general muscle stiffening, then rhythmic jerking of limbs. | Carbamazepine, lamotrigine, oxcarbazepine, valproic acid | Clobazam, lamotrigin, levetiracetam, valproic acid, topiramate | Gabapentin, phenytoin, pregabalin, tiagabine, vigabatrin |
| Tonic or atonic | Tonic: sudden general muscle stiffening, for $\sim$ 1 minute. Atonic: sudden loss of muscle tone. | Valproic acid | Lamotrigine | Carbamazepine, gabapentin, oxcarbazepine, pregabalin, tiagabine, vigabatrin |
| Absence | Seizure with arrest of current behaviour with EEG showing generalized spike wave activity. | Ethosuximide, lamotrigine, valproic acid | Clobazam, clonazepam, levetiracetam, topiramate, zonisamide | Carbamazepine, gabapentin, oxcarbazepine, phenytoin, pregabalin, tiagabine, vigabatrin |
| Myoclonic | Very short and sudden jerking movements. | Levetiracetam, valproic acid, topiramate | Clobazam, clonazepam, piracetam, zonisamide | see above (Absence) |
| Focal | Seizure is limited to one hemisphere, can be localized or widely distributed. | Carbamazepine, lamotrigin, levetiracetam, oxcarbazepine, valproic acid | Clobazam, gabapentin, topiramate | NA |
| Juvenile myoclonic epilepsy | Myoclonic seizure after waking, onset b/w 5-20 yo. Often also absence / generalized tonic-clonic seizures. | Lamotrigin, levetiracetam, topiramate, valproic acid | Clobazam, clonazepam, zonisamide | Carbamazepine, gabapentin, oxcarbazepine, phenytoin, pregabalin, tiagabine, vigabatrin |

#### 2.1.2.2 Surgery

Depending on the type of epilepsy syndrom, up to 70% of pharmacoresistant epilepsy patients can become seizure free after surgery (Baumgartner, 2001). Generally it can be said that surgery is a very effective treatment for refractory epilepsy or temporal lobe epilepsy. Further, some studies and surveys clearly confirmed improvement of quality of life after surgery in adults. Albeit, resective surgery is currently exclusively feasible for focal seizures, which originate from a specific source that can be clearly located in one brain hemisphere. The procedure of epilepsy surgery is mainly composed of three

gross steps. First, a very detailed pre-surgical evaluation has to be executed, to ensure that the respective patient is qualified for surgery, not responsive to any other treatment and doesn't fit the exclusion criteria. In a next step, the source of the seizure has to be located as detailed as possible through applying multiple brain research and imaging tools. Commonly, fMRI as well as a very detailed, often invasive EEG recording are used to identify the source. Again, being aware of the type of seizure the patient is suffering from is crucial for any further procedure. In a last step, the surgery is performed (Luders, 2008).

## 2.2 Electroencephalography (EEG)

For the purpose of this study data from electroencephalography (EEG) recordings will be evaluated. The following paragraph presents an overview of EEG as a brain research method and as a diagnostic measure for epilepsy patients.

### 2.2.1 Method and mechanisms

EEG enables researchers and physicians to measure the electrical activity of the brain. To monitor and record this activity non-invasively, electrodes are typically placed on the scalp according to the international standard 10-10 or the reduced 10-20 electrode placement system. Figure 4 shows all possible electrode positions according to the 10-10 standard electrode placement system, with the reduced 10-20 electrode positions highlighted in grey. Commonly a reduced number of electrodes is actually used when recording, depending on the individual need for high spatial resolution and detailed ability to assess the location of the neurophysiological origin of brain signals. In the clinical setting EEG signals are commonly recorded by a minimum of 21 electrodes (Zschocke & Kursawe, 2012). Further, if even higher resolution is necessary to locate brain lesions, identify epilepsy onset locations for surgery or detect oscillations of very high frequency, there is the possibility of invasive EEG systems, where electrodes get surgically implanted onto the brain surface.

The electrodes of the EEG pick up the electrical activity of postsynaptic potentials from large groups of pyramidal neurons in the upper layer of the cortex. In healthy humans at rest those cells mostly show asynchronous activity, while in many pathological conditions (e.g. epileptic seizures) groups of cells fire increasingly synchronous in the same frequency range. The activity of such neuron groups adds up and manifests itself in the form of higher amplitudes in the EEG recording.

### 2.2.2 Epileptic biomarkers in EEG

As already mentioned in Section 2.1.1, patients with suspected epilepsy are usually screened through recording routine EEG or if necessary even undergo the long-term proce-

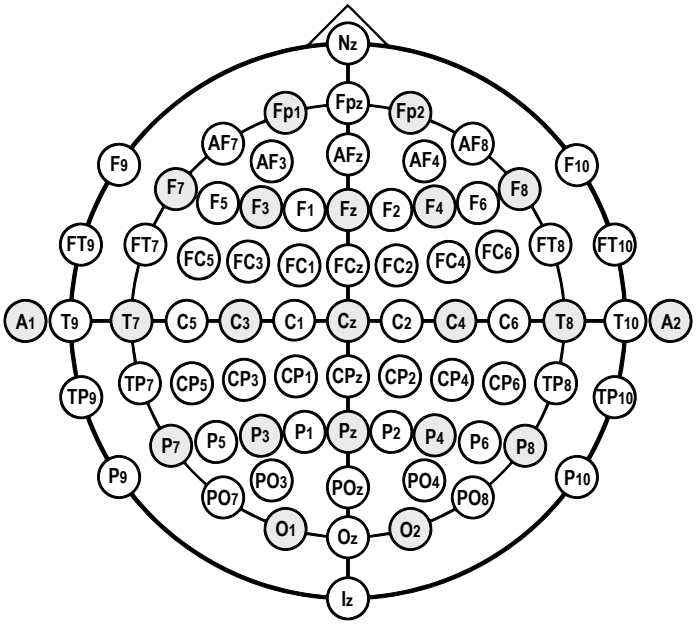


Figure 4: 10-10 electrode montage setup with 10-20 electrodes highlighted in grey.

dure of VEM. This measure is necessary to detect seizures and ideally estimate an approximate frequency of seizure occurrence. Seizures identified in a patient's EEG recordings are the most unambiguous biomarker of epilepsy, but as discussed above, the presence of such ictal events alone does not suffice to diagnose chronic epilepsy and predict future seizures. Furthermore, a correlation between interictal epileptiform discharges (IEDs) and epilepsy has been shown by several research studies and within recent years automated spike detection has been greatly improved. Nevertheless, spikes can also be found in otherwise healthy brain activity and many epileptic patients do not show spikes consistently, but only rarely or at specific times, e.g. at night. Identifying reliable electrophysiological biomarkers of epilepsy besides spikes and seizures could significantly decrease the monitoring times needed for a clear diagnosis and thus make faster appropriate treatment possible. Biomarkers which occur more frequently and consistently than seizures and spikes could therefore come very handy for the purpose of speeding up the diagnostic process. By now, such non-ictal biomarkers that can be utilized confidently for diagnosis have not been found (Engel et al., 2018). Though, meaningful interictal electrophysiological disturbances have been repeatedly found in routine EEG of epilepsy patients. The most reliable of such have been proven to be interictal spikes (IIS) and sharp waves, while invasive studies using wide bandwidth recording and small diameter electrodes have identified pathological high-frequency oscillations (pHFOs) and microseizures (Staba et al., 2014).

In the following paragraphs EEG snippets will be included for increased understanding of what is meant by epileptic EEG biomarkers in this thesis and how epileptic EEG differs

from normal, healthy brain activity. Those recordings were taken directly from the data sets used for creating the algorithm subject to this thesis.

#### 2.2.2.1 Seizures

In EEG recordings epileptic seizures are visually characterizable and quite clearly distinguishable from PNES (Baumgartner & Pirker, 2019). In Figure 5 a normal EEG recording is shown and compared to Figure 6 below, which shows a recording with seizure onset at approximately the second vertical line. The development of very rhythmic brain activity at seizure onset time followed by increased amplitudes is well visible in the center of the picture.

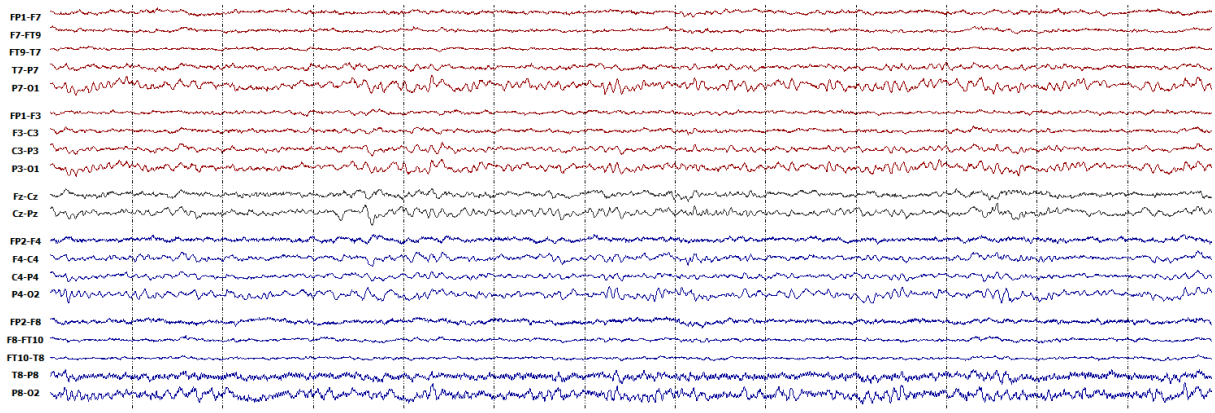


Figure 5: Normal, healthy-looking EEG recording during daytime.

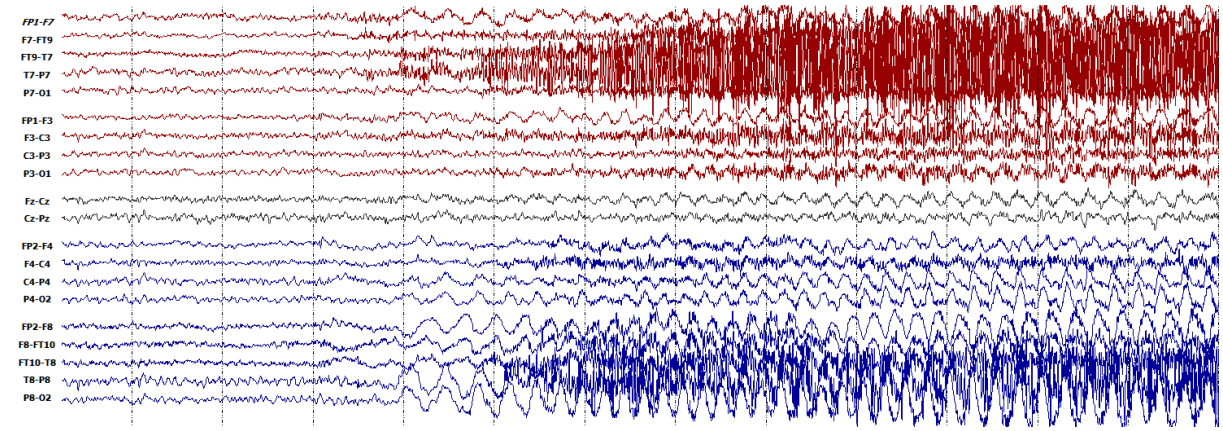


Figure 6: EEG recording of transition from normal activity to focal seizure.

This work will not focus on seizures and only include EEG recording snippets without seizure activity into algorithm training and testing. The goal here is to base diagnosis on different, more discreet biomarkers, that occur more consistently and might even be able to predict ictogenesis.

#### 2.2.2.2 Interictal epileptiform discharges (IEDs)

Interictal epileptiform discharges (IEDs) can show various different patterns, the most common and identifiable being spikes, spike bursts and sharp waves (Bagheri et al., 2019). Interictal spikes (IIDs) are characterized by a large-amplitude rapid component that is usually followed by a slow wave. IIDs can also appear irregular, regional, as bursts of spikes or combined with sharp waves depending on the particular epilepsy form. Spikes usually last between 50-100 ms, sharp waves consist of a rapid component lasting 100-300 ms and slow waves span over 200-500 ms (de Curtis et al., 2012).

In Figure 7 three examples of IEDs of different intensity and clarity are shown. All examples were taken from the data sets described in Section 3.1.

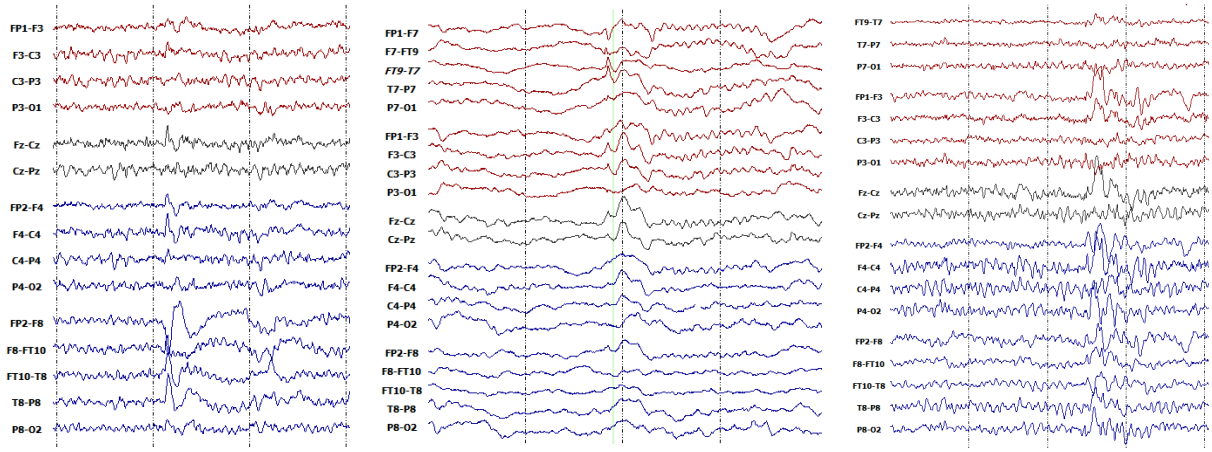


Figure 7: Comparison of two EEG snippets containing spikes. At the lower center of the left picture spikes are quite clearly visible, in the middle picture more subtle spiking behaviour is visible across channels right after the green line and the right picture shows a clearly visible polyspike wave complex.

Such IEDs and similar discharges are in the focus of the biomarker algorithm developed in this study. The goal of the proposed end-to-end algorithm is to identify such subtle interictal discharges in the data and deduce epilepsy risk from those detections. Additional underlying subtle patterns of IEDs are suspected and could help the algorithm to discriminate even better between epileptic and non-epileptic routine EEG recordings.

#### 2.2.2.3 Pathological high-frequency oscillations (pHFOs)

Pathological high-frequency oscillations (pHFOs) of 80+ Hz were first detected in the context of epilepsy surgery (Frauscher et al., 2017). Since then pHFOs have gained attention as one of the most informative biomarkers, with the ability to indicate the epileptogenic region of the cortex and potentially even to indicate epileptogenesis and ictogenesis (Engel et al., 2018). Unfortunately a large proportion of previously recorded EEG data has most likely not been recorded with sufficient resolution to be able to consistently and realiably

detect such pHFOs. Thus, even though pHFO detection will be subject of further research within this project, it will not be targeted in particular in this thesis.

#### 2.2.2.4 Theta rhythm reduction

Epileptogenic IIDs seem to reduce theta rhythm (4–7 Hz frequency range) in the brain. Research suggests that theta reduction by up to 60% happens right after each spike in all patients (Fu et al., 2018). In some patients, especially such without a clear epilepsy onset location or several focal points, Theta rhythm reduction extends to between-spike periods.

#### 2.2.2.5 Connectivity

Research in functional neuroimaging has increasingly focused on connectivity in recent years and some astonishing insights were found by those means. This approach targets explanation of inter-communications and pathways in the brain instead of focusing on dedicated brain regions for specific tasks. Studying the connectivity of the epileptic brain seems to be promising for seizure prediction and epileptogenic focus localization (van Mierlo et al., 2014). If an interictal EEG biomarker based on connectivity could be detected, this could potentially be a robust and reliable way of predicting epilepsy from relatively short EEG recordings. This means of analysis has already been pursued by various researchers in the field of epilepsy (Centeno & Carmichael, 2014). Such connectivity analyses combined with machine learning methods as applied in this study can offer interesting insights. Diving deeper into that opportunity unfortunately exceeds the scope of this thesis.

## 2.3 Artificial Intelligence

This section explores the theoretical framework of the computational methods applied in the study at hand. Starting with a conceptual description of artificial intelligence in general and convolutional neural networks and residual networks in specific, this chapter will go on to discuss the current state of the art in applying AI in healthcare, where ethical considerations on AI assisted medical diagnostics will be taken into account. A subsection is dedicated to the theory and implementation of explainability in deep learning algorithms and explores possibilities to ensure the development of trustworthy systems.

Science and technology started fostering the idea of automatic, self-regulating control under the term Cybernetics back in 1948. The term *Artificial Intelligence* was coined in 1956 and helped to spark the public debate about this concept (Russell, 2014). Since then, this comparatively new field has gone through periods of varying popularity, closely tied to ongoing discussions about its potential and associated risks. Within the last decade,

the power of data has gained importance globally. With more focus on data science and immensely increased availability of data in various domains, AI has simultaneously gained popularity again. The reasons for this were first and foremost the tremendously increased capabilities of AI methods given increased amounts of data to learn from. To understand those concepts more thoroughly, the notions *Artificial Intelligence* and *Machine Learning* will be defined and some technical assets will be highlighted in the following. Figure 8 visualises how the AI concepts discussed in this thesis relate to each other.

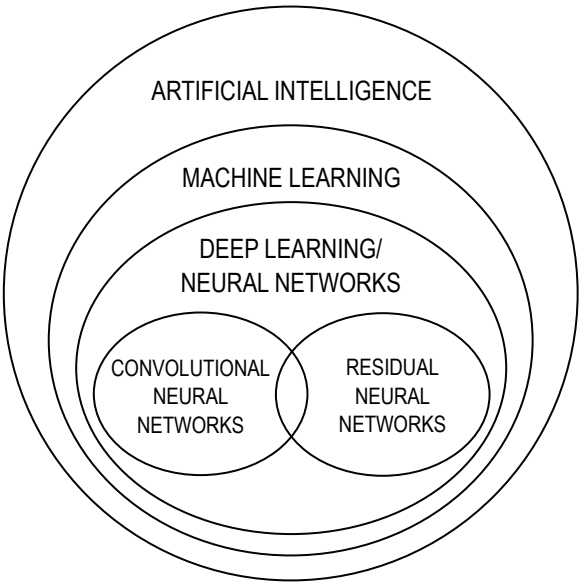


Figure 8: Visual definition of how the discussed AI concepts interrelate.

Artificial intelligence is the broader, more general concept, while machine learning strictly constitutes a subcategory of AI. However, currently applied AI technically can be considered almost identical with ML. While future innovations might enable other ways to create some sort of intelligent behaviour in machines, at this stage of AI advancement most methods in use and under development fall under the umbrella of ML.

The term ML denotes computational models and algorithms that are designed to learn from input data, similar to how humans are believed to learn from experience and update their beliefs about the world. In other words, ML algorithms are programmed in a way that enables them to automatically draw conclusions from 'training data' that generalize to an independent set of 'test data'. Those predictions and decisions drawn from data are made and acted upon without being explicitly programmed to do so, but instead for every new 'experience', internal computations are performed and learning parameters are updated.

Using the human brain as a model for developing learning algorithms was a big step towards more efficient systems. Artificial neural networks (ANNs) similar to the ones developed nowadays were first built back in the 1980s. Even though they were already

functioning similarly as we know them today, their actual power was only made visible in recent years due to the increasing availability of massive amounts of data and increased computing power of GPUs. Input is processed by various nodes and huge amounts of different connections are constantly updated until a satisfying classification accuracy is achieved. Subsequently the network is able to correctly classify and evaluate novel incoming data. However, just as in human mundane decision making in day to day life, it is usually hardly comprehensible for an outside spectator what exactly led to the decision at hand. And while humans might be able to explain the process which led to making that decision, algorithms usually fail to do so.

### 2.3.1 Artificial Neural Networks

Deep learning is a subcategory of machine learning and usually refers to methods based on artificial neural networks (Goodfellow et al., 2016). Those systems are built using a deep architecture which consists of several layers of operations between input and output. This can on a high level be compared to its role model, the brain. Each layer of the ANN consists of so-called neurons, which pass on a signal by performing an operation and sending the weighted result on to the next neuron. Through this complex architecture such models are able to learn high-level patterns and hidden features in the input data. This technology can be applied to classification as well as regression problems and very diverse data types. This fundamental idea has been further developed into countless variations and different approaches depending on the problem and the input data. In the following, the conceptual background of the two ANN variations used in the project at hand will be described. The concrete implementation of those will be discussed in the methods Chapter 3 of this thesis.

#### 2.3.1.1 Convolutional Neural Networks

Loosely inspired by the brains visual cortex, modern convolutional neural networks were first developed in the 1990s. Accordingly, CNNs were originally mainly applied to image recognition challenges. Within the last years of research and development, implementing convolution operations into ANNs has been proven to be a great asset also for audio, language and time series classification. A CNN learns so-called feature filters during training, which enables it to detect e.g. edges and shapes in images. Depending on the input data, the learned features can correspond to different types of data patterns. Using this technology on time series, and more specifically on EEG data, it can be expected to detect typical shapes like IEDs and thus learn to classify the input data accordingly. What sets this technology apart from other ANNs is its usage of the convolution operation. This mathematical operation is an element-wise multiplication and addition as depicted in equation 1.

$$s(t) = (x * w)(t) \tag{1}$$

In the above equation, $x$ represents the input and $w$ the so-called kernel. In the context of CNNs, a kernel is an array storing the weights that are adapted during training. Depending on the problem formulation and the input data, several kernels can be stacked and used as a bigger filter. The filter size is usually chosen considerably smaller than the input size, so that the convolution is applied to the input step by step when sliding the filter over the input data. The smaller the filter size is compared to the input, the more detailed features will be detectable. Convolving the kernel with an input over an index (e.g. EEG recording time $t$) results in a so-called feature map ($s(t)$). With every convolutional layer applied in a CNN, more detailed feature maps are created (Goodfellow et al., 2016). As this process is hard to image from pure verbal description, please refer to the visual explanation of a 2D convolution in Figure 9.

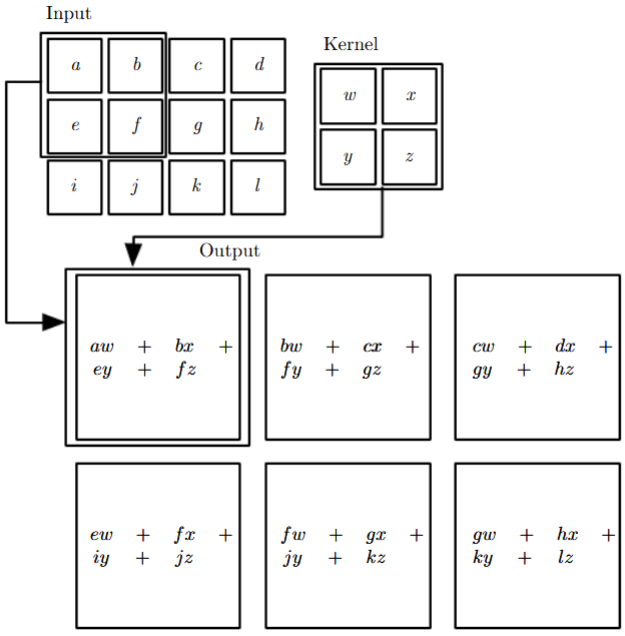


Figure 9: 2-D convolution example by Goodfellow et al. (2016).

After each convolutional layer commonly a pooling layer is applied to compress the information. This is usually done by sliding another filter over the convolution output. This filter either chooses the maximum or averages over the output of each piece it sees. Whether average or max pooling is more appropriate depends strongly on the problem formulation and the input data. After all convolutional and pooling layers are performed, usually fully-connected layers, or also called dense layers, are applied. Here all neurons are connected to all output options and a classification or regression result is created. As in any other supervised learning algorithms, at this point during training the result

created by the network is compared to the ground truth or label of the input data and the error is back-propagated to update the weights of the network.

#### 2.3.1.2 Residual Networks

Another subtype of artificial neural networks are residual networks (ResNets), which consist of so-called residual blocks. This technology was built to resemble a similar structure as pyramidal cells in the human cerebral cortex. Hence, it is implemented in a way that connections between layers are skipped through so-called 'shortcuts'. Those shortcuts are added at the end of each residual block and usually either are identical to the input of the corresponding block (identity block) or one simple convolution operation is performed on the shortcut (convolutional block). Figure 10 depicts the mechanisms of residual identity and convolutional blocks in the manner they were applied in the algorithm presented in this thesis. This type of network can have a few proven advantages over classic plain networks. First of all, it makes building deeper networks possible, while at the same time residual connections often get rid of the vanishing and exploding gradient problems (He et al., 2016).

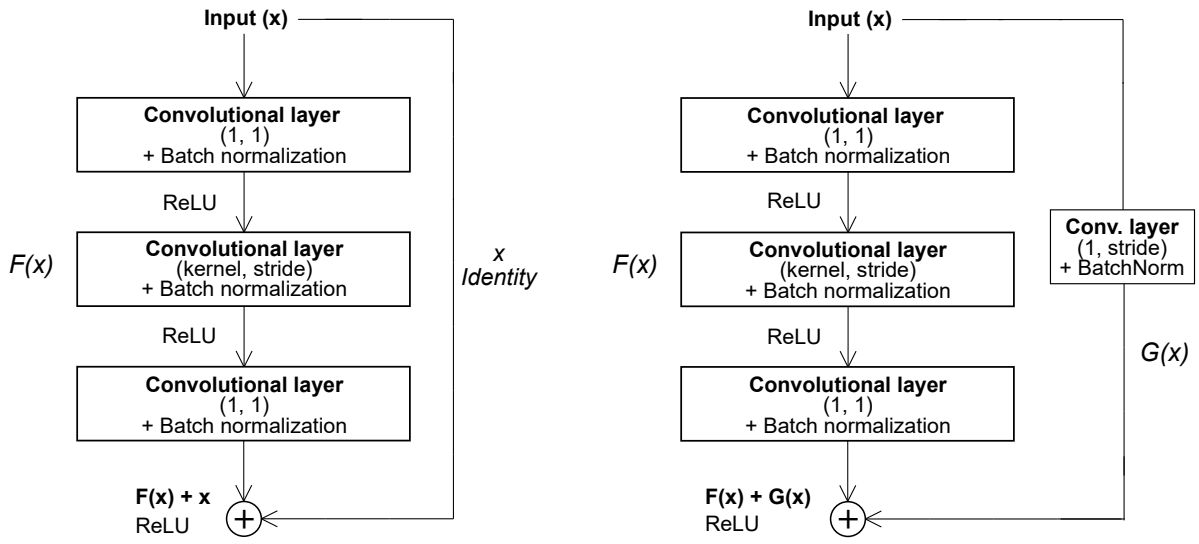


Figure 10: Residual identity block (left) and a shortcut convolutional block (right).

### 2.3.2 Deep learning in medical diagnostics

There have been recent studies using deep CNNs on EEG recordings in a similar manner as the study at hand. Researchers in several other studies have trained a CNN to distinguish normal from abnormal EEG recordings with over 80% accuracy (van Leeuwen et al., 2019). IED detection algorithms based on deep learning methods have also been

developed (Fürbass et al., 2020). Further, the implementation of deep learning for predicting brain maturation from EEG recordings of premature neonates showed promising results (Gschwandtner, 2020).

Applying novel AI methods to the field of medicine brings along ethical challenges concerning trust and transparency of such systems in clinical use. At the current fast pace of development in the field of AI, evaluating measures to make consistent ethical decisions constitutes a very important foundation. AI researchers and developers should be aware of ethical aspects of their work and integration of potential solutions should be self-evident. In the following sections those considerations and the current state of the art in explainable AI will be elaborated on in more detail.

#### 2.3.2.1 Explainability and transparency

Applying AI and ML methods to issues in the medical field and especially to medical diagnostics and decision support currently appears to be a popular topic. Besides many successes of medical AI, this topic carries ethical conflicts and issues to debate.

Due to its technical characteristics and functioning, ML usually does not offer a straightforward way to understand its reasoning. Specifically the technology of deep ANNs is often referred to as a 'black box', as the procedure of how such algorithms reach their conclusions are hardly explainable. As described in Section 2.3.1, deep ANNs are based on several layers of computations creating great amounts of parameters and high-level features which are not explicitly coded but automatically deduced from input data. Thus, after numerous layers of computations a clear meaning and origin of each of the obtained features is not transparent and comprehensible anymore.

This lack of explainability makes it often rightfully difficult for medical personnel to trust in systems based on ML technologies. Furthermore, this characteristic of deep neural networks can be accidentally or deliberately exploited through feeding incomplete or biased data to the system during training. An ANN algorithm strictly learns what it is taught by the data. Hence, ML systems for medical use should be handled with great caution and only be applied after careful consideration and under supervision of informed medical professionals. Before trusting a system to an extent where diagnosis and treatment is based on its evaluation, the algorithm design and the data used for training should be investigated and understood to prevent the integration of fatal biases. For example, a model accidentally trained only on data of male patients could be biased towards this data type and overlook the disease in females if the manifestation and symptoms are slightly different. Issues like this are fall under the term *gender data gap* and are fortunately increasingly debated recently.

However, inspecting not only the composition of the input data itself but also which features of this data the algorithm is actually paying attention to poses another key

problem. Understanding what exactly happens within the hidden layers of a deep ANN is not even accessible to the ones building it. ML engineers thus have a hard time optimizing their ANNs deliberately when the results do not turn out as expected and they have no clear evidence helping them to understand why. Currently a lot of heuristics are involved when optimizing ML models, parameters are tweaked and layers are added in the hope of more or less coincidentally creating better accuracy. This process can be much more straight forward with more insight into the workings of the algorithms. Therefore, growing numbers of analysis methods for machine learning algorithms are being developed and the usage of such tools is gaining popularity. The importance of creating interpretable and explainatory AI is increasingly being highlighted from various angles and promoted by scientists in the field. Even though literature mostly concludes that current methods are insufficient, new opportunities and ethical ideals are continuously discussed, as this will remain a crucial issue as long as AI methods persist (Gilpin et al., 2019).

Before the available methods are elucidated in more detail, a definition will be given of what is meant by the notions *explanation* and *understanding* in the context of this thesis. According to previous literature (Montavon et al., 2018), the term *understanding* talks about functional understanding of the networks decision making behaviour, not about understanding algorithmic details of its computations. Hereby the aim is to understand individual decisions of the model. On the other hand, an *explanation* of a model decision provides several human-comprehensible features which have presumably contributed to the respective decision. For instance, a DNNs prediction of high epilepsy risk in a specific patient can be understood, if the explanation is provided that there were several spikes, HFOs and a reduced theta rhythm in the respective recording of that patient.

A notable amount of different options for achieving explainable AI systems has been developed over recent years. Table 2 shows an overview over the most popular modern methods. Explainability methods can be divided into three different types depending on their strategy and point of implementation. Thus, there are pre-modeling, modeling and post-modeling approaches (Fellous et al., 2019).

*Pre-modeling explainability* mainly deals with the input data before training the model. Input data can be characterized, feature extraction can be performed and more. However, this option inherently involves the risk of biasing the model. Another possible side effect is decreased model performance due to too many set defaults and not letting the model learn potentially unexpected or unknown features. If this is done unrestrictedly anyways after pre-modeling measures were taken, we face the risk of the model ignoring those pre-modeling results, which means that this pre-modeling evaluation actually does not explain what the model bases its decisions on. *Modeling explainability* includes all methods integrated into the model itself. This usually means that during training an explanation gets trained as well in some way. In the end the model is able to explain its workings 'itself' while it is applied to new data samples. Thus this variant can be described as the

Table 2: Approaches to explainable AI systems, adapted from Fellous et al. (2019).

| Type | Pre-modeling | Modeling | Post-modeling |
|---|---|---|---|
| **Goal** | Characterize input data | Design explainable model architectures and algorithms | Extract explanations from outputs |
| **Methods** | **1. Exploring data analysis:** beyond reporting statistical properties. **2. Data set description & standardization:** describing variables, metadata, provenance, statistics, between variables (pair plots, heatmaps), ground truth correlations, probabilistic models generating synthetic data. **3. Explainable feature engineering:** Using the right features to explain predictions is crucial. **4. Data set summarization:** Interpretable prototype selections and identification of meaningful outliers. | **1. Adopting a more explainable model family:** linear models, decision trees, rule sets, generalized additive models, etc. **2. Hybrid explainable models:** Deep k-Nearest Neighbors (DkNN), Deep Weighted Averaging Classifier (DWAC), Self-Explaining Neural Network (SENN), Contextual Explanation Networks (CEN), Bag-of-feature network (BagNets). **3. Joint prediction & explanation:** Teaching Explanations for Decisions (TED), Multimodal/Visual explanation, Rationalizing Neural Prediction. **4. Explainability through architectural adjustments:** Explainable CNNs, 'This Looks Like That' architecture, Attention-based models. **5. Explainability through regularization:** Tree Regularization, Reg. by local explanations constraints. | **1. Explanation targets:** Mechanistic (internal, algorithmic) vs. functional (external, interpretative) explanations at different complexity levels. **2. Input based explanation drivers:** How input manipulations can drive/change the outputs. **3. Macro-explanations:** Importance scores, Decision rules, Decision trees, Dependency plots, Verbal/Counterfactual explanations. **4. Explanation estimation methods:** Perturbation-based (LIME, SHAP), Backward propagation, Proxy model, Activation optimization. **5. Be careful with:** manufacturing explanations or over-interpreting the outputs! |

most transparent. In this method one does not project anything on the learning process before or afterwards, but information is extracted directly from the process. Unfortunately though, implementing such explanation mechanisms right into the model itself in practice often turns out to be the most complicated option. In *post-modeling explainability*, information about the network and its decision making is extracted from the output. This can for example be done by using decision trees or other methods to reconstruct the decision making process given the output choices. Conceptually such methods can give the impression of 'guessing' why the network potentially decided the way it did. As with pre-modeling, general causality is the main target here. This approach makes it possible to create great results for some problem formulations, but the actual transparency is debatable. Also post-modeling is mostly useful to more or less artificially create a user-friendly explanation in the end, but doesn't help engineers developing such algorithms to

figure out what goes wrong within their network.

#### 2.3.2.2 Precision medicine

With precision medicine another healthcare trend emerged within recent years and attracted attention from different fields and perspectives. Epilepsy is a particularly diverse disease, where personalized factors are essential for diagnosis and treatment. Hereby genetic testing has gained importance throughout recent years (Kearney et al., 2019). Besides novel genetic approaches, utilizing AI methods to personalize diagnosis and treatment can further offers many novel opportunities to improve the current efficiency of medical care around the world. In the following, this thesis project will be tackled with the values of precision medicine in mind. This project attempts to position itself within the scope of precision medicine to identify additional challenges revolving around the epilepsy syndromes.

# 3 Methods

This chapter summarizes the methods used in the research project presented in this thesis. The data used for training and testing the proposed neural network are introduced. Further, this chapter explains the designed algorithms, the experiments performed with those models and the statistics used to evaluate their performance.

## 3.1 Data sets

For this research project, EEG data from six different sources were available, comprising a data set with routine EEG recordings of 3568 patients in total. 3345 EEG recordings of the publicly available EEG corpus provided by the Temple University were included, while 475 of those were taken from the seizure subset and 2870 from the normal/abnormal subset. 100 recordings with diverse diagnostic outcome were provided by NRZ Rosenhügel. Recordings of 31 patients with clear epilepsy diagnosis were provided by the University Clinic Erlangen, 18 by Kempenhaeghe epilepsy center and 14 by Vienna General Hospital (AKH). Additional recordings of another 60 patients (30 epilepsy, 30 non-epilepsy) were provided by the Danish Epilepsy Center Dianalund. For the purpose of training and testing the proposed end-to-end neural network not all of those recordings were suitable. The data sets were split up into clearly annotated data and recordings without a diagnosis or annotation of spikes. The latter could not be used for now in this supervised training path, as there is no ground truth available. Thus, for the purpose of training this network to differentiate epileptic EEG from non-epileptic EEG, the Erlangen, AKH, Dianalund and Kempenhaeghe data sets and 56 spike and 31 control recordings from the NRZ data set were used. Additionally the medical documentation text files were scanned for all recordings and those with clear spiking behaviour and an epilepsy diagnosis were annotated. Those data, as well as TUH recordings annotated as being seizure and spike free without a history of epilepsy were included in the end-to-end training and testing. This results in recordings of 387 patients (241 epilepsy and 146 non-epilepsy) used in most of the experiments described in this thesis. Later, 116 additional epilepsy patients without any epileptiform EEG activity were included in a subset of the performed experiments.

For most experiments the Dianalund data set was excluded from training and validation to serve as an independent test set in the end. This choice was made, because of the size and the balanced class distribution of the Dianalund set. However, subsequently additional experiments were performed with a random balanced independent test set. For this, 60 patients (30 epilepsy, 30 non-epilepsy) were chosen at random from all 387 patients and removed from the data set to serve as a final test set. The remaining 327 patients were split into training and validation set with a ratio of 90/10, thus 294 patients were used for training and 33 for validation. The distribution of the data sets

used in most experiments can be seen in Figure 11. To compensate for the imbalance in the training set the experiments were performed with oversampling and undersampling of the data set respectively to evaluate the more effective technique for the particular use case at hand. The data were pre-processed by applying a low-pass filter at 49 Hz and then downsampling to 128 Hz. For further experiments the data were pre-processed using the tool PureEEG for automatic artifact reduction (Hartmann et al., 2014).

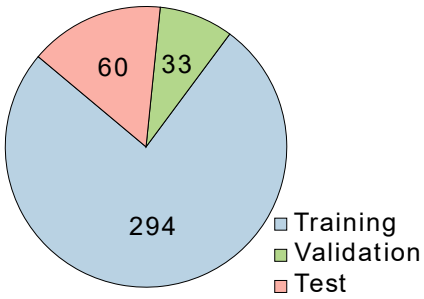


Figure 11: Distribution of the data into train, validation and test set.

## 3.2 Algorithm design

The focus of this master's thesis was to develop an end-to-end algorithm, aiming to identify epilepsy biomarkers in raw EEG data and classify the recordings accordingly. Figure 12 provides a schematic graphic description of the general network structure used within the presented project.

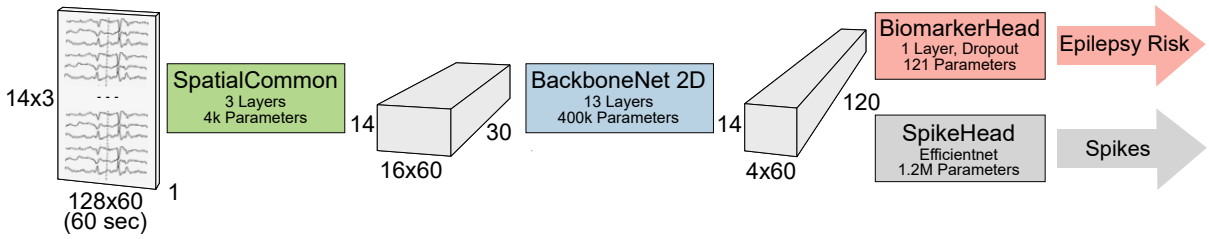


Figure 12: Overview over the algorithms used in the entire project and their relationship.

In all applications of this network structure the data is first sent through a few layers of computations with the purpose of handling the spatial information contained in the data, which is referred to as SpatialCommon network in the following. Hereby, two convolutional and two pooling layers were used for extracting features from the spatial dimension of the EEG data and then reducing the redundant dimension. Next, the spatially reduced data is fed into the BackboneNet, which is in the focus of this thesis. The results obtained from this network are subsequently classified through a fully-connected layer (BiomarkerHead) and the estimated risk for epilepsy is returned. At this point it is possible to train the

additional 'SpikeHead' in parallel, which allows the network to specifically learn IED patterns from accordingly labelled data. However, the spike training is not part of this thesis. Instead, the presented project focuses on general end-to-end training using the BiomarkerHead, with the aim to classify the data based on any salient patterns that can be found consistently.

In the following, all network architectures that were experimented with in this study regarding the end-to-end training will be presented. The algorithms were built using Python and the deep learning framework TensorFlow by Google Brain (Abadi et al., 2015). The training process utilized Adam (Adaptive Moment Estimation) optimizer. Each convolutional layer is followed by a ReLU (rectified linear unit) activation function and subsequently batch normalization is applied. Besides batch normalization, further means for regularization were implemented. To account for the risk of overfitting, most networks presented in the following include at least one dropout layer in there architecture. Further, the training data was slightly augmented by adding random noise and by applying amplitude alteration (random between 0.8 and 1.2) and spatial flipping.

### 3.2.1 Basic Convolutional Neural Network

At first a rather simple model was trained, before complexity was steadily increased with the aim to optimize performance. First the network consisted of 3 convolutional layers (7×1, 3×3, 3×1), with a maximum pooling after each layer and another final maximum pooling layer, located after reducing the dimensions and before the dense layer. In a next step the kernel was changed to (5×5, 3×3, 3×1) and a dropout (0.5) was added. The third step here was to change the last pooling layer from maximum to average pooling. All those indicated steps positively affected performance and accuracy of the model. The final version of this network is depicted in Figure 13. Here also the SpatialCommon network mentioned above is visualized in detail. This structure for handling spatial information in the data is implemented in the same manner in all networks used in this project.

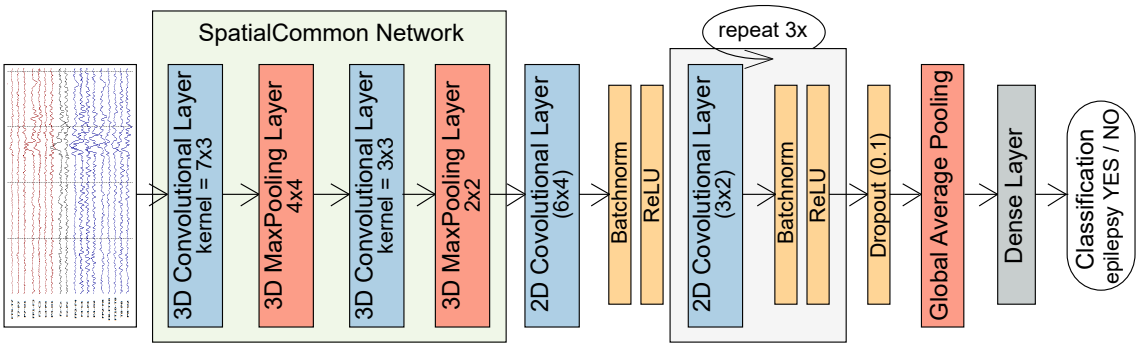


Figure 13: Basic convolutional neural network (BasicConvNet).

To further increase depth and number of feature maps, the next step was to implement a residual structure.

### 3.2.2 Residual Network

Inspired by comparable previous research and the current trend in the ML community, a residual network structure was chosen. This structure promises faster convergence despite the possibility of greatly increased depth. A recent study by Lu & Triesch (2019) successfully utilized a residual network structure for EEG signal classification in epilepsy. They showed promising results for applying a network consisting of two iterations of one convolutional block and two identity blocks each.

For building a residual network architecture for the project at hand, increased complexity had to be considered due to the data not including seizures and the specific goal to differentiate short non-ictal recordings. In Figure 14 the final structure of the basic residual end-to-end biomarker network is shown and in Figure 15 its structure with the use of attention pooling is depicted.

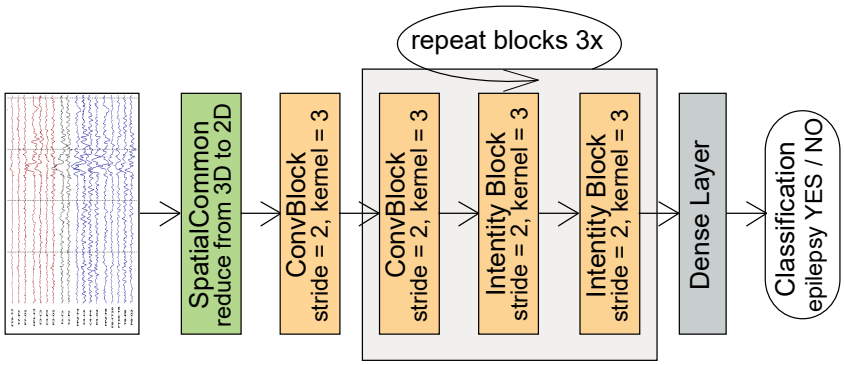


Figure 14: Residual network for end-to-end training (BasicResNet).

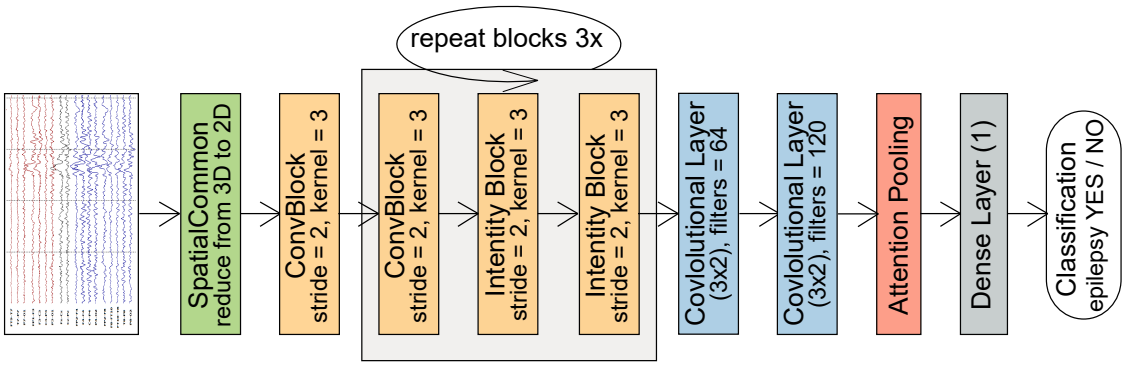


Figure 15: Residual network with Attention Pooling (AttentionResNet).

Likewise as in all neural networks built during this project, the data is first sent through the simple SpatialCommon network (the detailed structure was shown in Figure 13). Then the spatially reduced data is fed into the BackboneNet, which in this case is built up by three repetitions of the residual structure. Each repetition consists of one convolutional and two identity blocks. The structure of the residual blocks was described in detail in Section 2.3.1.2 and depicted in Figure 10.

For the purpose of creating an intrinsically explainable algorithm, another network structure was built. In this case, a novel technique called attention pooling is applied (Trinh et al., 2019), which will be explained in more detail in Section 3.2.3 below. Several experiments regarding this network architecture were performed. One larger variation of the AttentionResNet included a second dense layer with 120 features in the end and a dropout layer between the two dense layers, while another, more reduced version went from the residual blocks right into the attention pooling layer and did not compute any further layers except for the final dense layer.

### 3.2.3 Explainability

In this section all considered or attempted methods to create a more interpretable, explainable CNN will be reviewed and discussed. This ranges from approaches implemented right into the network (ProtoPNet, Attention) to post-hoc methods (SHAP, LRP). The four techniques mentioned were chosen to be investigated in detail and compared. The aim was to compare the different approaches and see which one offers the most trustworthy and fast results and is straightforward to implement.

#### 3.2.3.1 ProtoPNet

The first idea regarding a suitable explainable AI technique in this project was an adapted version of the ProtoPNet (C. Chen et al., 2018). ProtoPNet is designed to be trained using an explainable network architecture right from the start. During training the model learns to identify so-called prototypes for each class. For every novel inference to the trained model, it maps the input to the most similar prototype. The prototypes can be plotted in the end with the classification results for comparison. In the original application a big data set was used consisting of pictures of birds classifiable into 200 different bird species.

For several reasons the task at hand is quite different from the original use case. First of all, this thesis project deals with a binary classification problem (epilepsy vs. no-epilepsy). Additionally, of those two classes only epileptic EEG shows a distinct structure, while non-epileptic EEG can be a recording of a patient suffering from any other disease but epilepsy and can therefore include very diverse patterns. Accordingly, it will be quite difficult to define a reasonable prototype of non-epileptic EEG. Another challenge here is the fundamental difference in input data. Even though in CNNs EEG recordings are

treated very much like images, those data types are not much alike and their dimensions and features differ greatly.

#### 3.2.3.2 SHAP

When researching in the area of explainable AI tools, LIME is one of the XAI techniques that pops up first. It was developed with the aim to be capable of creating explanations for any classifier (Ribeiro et al., 2016a). The algorithm is treated as a black box here. Thus, the created explanation is not based on the actual workings of the algorithm, but is deduced post-hoc. Even though this approach is not ideal for the medical use case at hand, evaluating it and comparing the results seemed reasonable. As LIME was complicated to install in the specific configuration used in the project (Python 3.8, tensorflow 2.2.0, etc.), finally the decision fell on SHAP (SHapley Additive exPlanations). The basic idea is similar, however in SHAP a game theoretic approach is taken. Shapley values are computed to evaluate the contribution of each involved high-level feature. Hereby the features have to be defined beforehand. Those Shapley values are then used to reason and explain the networks decisions (Lundberg & Lee, 2017).

#### 3.2.3.3 Attention pooling

This approach to interpretable deep learning is implemented right into the neural network itself. Hereby, attention mechanisms are utilized (Vaswani et al., 2017) and applied in a novel manner in form of attention pooling (AP), similar to how it was proposed by Trinh et al. (2019). Attention learns to focus on relevant parts of the input to reach the optimization goal. In this work a transformer layer originally used in natural language processing (NLP) (Vaswani et al., 2017) is utilized. Several attention layers can be applied within a network, for example to focus on important temporal as well as spatial information. Considering the example of applying attention over time, the attention layer performs a dot product to detect similarities of previously successful detected input patterns to minimize the loss. In attention pooling (AP), these attention transformer layers are used to perform pooling. The network is taught to compress information of sequence of input vectors into a single vector. By performing those computations, the attention pooling layer enables the network to learn how to compress a sequence to a single element with minimal loss of important information. As opposed to classic convolutions basically detecting individual patterns in data, this type of learning is able to efficiently recognize important patterns in long sequences of data that resemble sequences seen before and match them to each other. Recognizing this correspondence is possible even though context and patterns of sequences are not identical. This mechanism can be very useful in NLP for detecting specific words succeeding others, no matter how many filler words would occur in between. In the use case at hand, the idea is that some IEDs might be

meaningful if they occur in a specific sequence. Attention pooling could learn and point to such IED patterns. These calculations subsequently return the so-called attention vector the network has learned. The values of this vector can be plotted on the input data to visualize what the network is paying attention to.

In this project, one AP layer replaces the usual max or average pooling after the last convolution. Here the attention layer (tf.keras.layers.Attention) precedes the final, fully-connected dense layer in the very end of the network (Figure 15. This is convenient, because at that point the temporal resolution of the data is available to be processed by the attention pooling layer. Thus, here the attention vector is only able to highlight a point in time and not specific patterns located on single electrodes. This is due to the performed reduction of spatial dimensions in the beginning (SpatialCommon). For additional spatial attention another attention pooling layer would be necessary earlier and more spatial features need to be computed before this layer. Integrating and optimizing more than one attention pooling for spatial resolution exceeded the scope of this thesis, but is planned to be explored in ongoing reseach.

#### 3.2.3.4 Layerwise relevance propagation

Another approach that is increasingly popular and seemed well suitable was the so-called layerwise relevance propagation (LRP), proposed by Lapuschkin et al. (2016). This approach analyzes the model outcome by propagating a specific computation backwards through each layer and then creates a heatmap for visualization of the explanation. Hereby various propagation rules can be applied, depending on the model and the desired complexity of the rule. The different rules arise from specific choices of the parameters $\alpha$ and $\beta$ in equation 2. The simplest LRP propagation rule with $\alpha$=1 and $\beta$=0, also applied as deep taylor decomposition, is expressed in equation 3.

$$R_j = \sum_k (\alpha \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} - \beta \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-}) R_k \tag{2}$$

$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k \tag{3}$$

The computed relevance scores resemble to what extent each feature contributes to the networks decision, considering the actual feature maps computed by the given network. A vector of the same size as the input is created from those relevance scores, which for explanation purposes can be plotted in the form of a heatmap. Heatmapping highlights the parts of the input data which were the most and the least relevant for creating a classification. With higher order LRP rules ($\beta \neq 0$) in use, the parts (e.g. pixels or time points) of the input data that contribute to the prediction (positive relevance) are high-

lighted in red and parts that contradict the prediction (negative relevance) are highlighted in blue. Using the simplest rule only the red highlighting is applied.

Unfortunately, previous literature suggests, that LRP heatmapping is associated with some systematic artifacts if applied to CNNs with residual structure (Rojas et al., 2019). Therefore, for this thesis project a specific SimpleConv architecture was created with the aim to be used to simulate LRP heatmapping.

## 3.3 Experiments

Various different experiments were performed, using different network architectures and varying hyper-parameters and input compositions. A high-performing subset of those trials will be described in the subsequent section. The residual network structure was expected to yield promising results for the end-to-end network, so most further advancement of this project focused on optimizing the ResNet architecture.

### 3.3.1 Network architecture

First of all the architecture itself, namely the number of convolutions, size and type of pooling and the number of residual blocks, was modified.

The residual network was first kept quite shallow, consisting of just three residual blocks, each consisting of three convolutional layers and a residual (shortcut connection) added in the end. This structure expanded up until a depth of one simple convolution followed by four iterations of three blocks each. Including the spatial layers at the start, this deepest setup added up to 39 convolutional layers.

The simple CNN architecture was continuously adapted and evaluated too. Hereby experiments were performed with a network depth ranging from five to ten convolutional layers. Again spatial dimension were convoluted and reduced first, but also here the number of layers was varied.

### 3.3.2 Loss function

Next, different possibilities for effective loss functions were compared. Mean squared error (MSE), Root mean squared error (RMSE), Mean absolute error (MAE), Squared error (SE) and Tukey Loss (biweight or bisquare function) came into use. Tukey function is known to be more robust and has notable benefits when trying to eliminate larger errors without being too sensitive to outliers. The function behaves rapidly around zero and only marginal changes for errors exceeding a variable threshold (see Fig. 16).

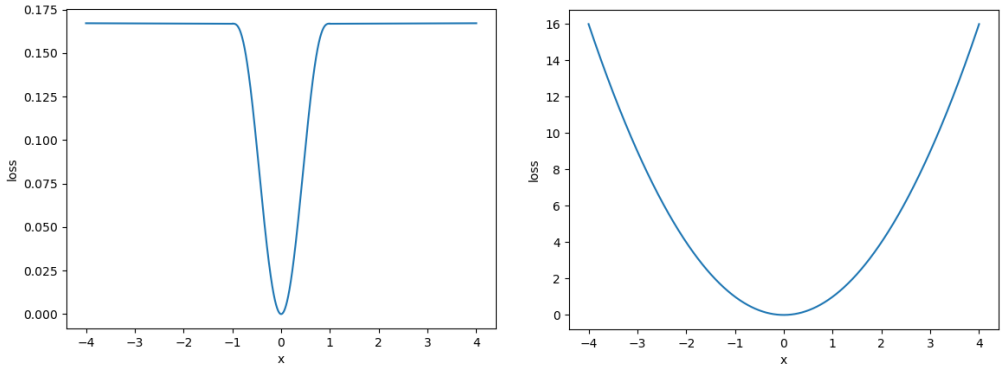


Figure 16: Comparing Tukey (left) and Squared Error (right) loss functions.

### 3.3.3 Hyper-parameters

Lastly and most importantly many parameters were adaptable and experimented with in diverse combinations. The size of compared learning rates ranged between 0.2 to 0.00001 and a large variety of options was experimented with. In all experiments learning rate warm-up and exponential learning rate decay were applied to adapt the learning rate over time. Hereby different decay rates and step sizes were experimented with. Also different training batch sizes were examined, from 8 up to 32 patients per batch. Also the input features were varied between 8, 12 and 16 features. Per residual block the features were increased by the factors 1.5 up to 9.0 depending on the depth of the network with steps of 1.5.

### 3.3.4 Input data

The same recordings of non-epileptics were used in all experiments. All of them were diagnosed with not having epilepsy and showed normal EEG activity.

Regarding the epilepsy data sets, several setups were experimented with. At first the training data only consisted of epilepsy recordings without seizures but including at least one visible spike. This was motivated by the assumption that it would make it easier for the network to recognize a difference between the sets. Next, recordings of epileptics which do not show any spikes nor seizures were included additionally. Finally, experiments were performed by training the network only on those no-spike recordings.

Also the impact of pre-processing and artifact reduction in the input data was investigated. For this experiment all non-epileptic recordings and the initial epilepsy set (no seizure, visible spikes) were pre-processed using the automatic artifact reduction PureEEG (Hartmann et al., 2014).

Furthermore, experiments on the class distribution of the training and validation data sets were performed. As the used data consisted of more EEG recordings of diagnosed epilepsy patients than of non-epileptics, this imbalance had to be made up for when

training the networks. In oversampling experiments the non-epileptic recordings included in the training set were duplicated until both classes reached approximately the same number of training samples and the distribution of positive and negative labels in the data yielded about 50%. For undersampling, the training recordings of the epilepsy class were reduced by a factor that again resulted in the two classes reaching approximately 50% of distribution across the training set. Hereby the necessary number of recordings was randomly chosen and then removed from the data set.

Another experiment looked at the distribution of the data in the training, validation and test set. As already mentioned above, the validation set was always directly chosen at random from the training set. The test set, on the other hand, needed to be independent and was only applied after training was completed. In a first trial, the test set only consisted of recordings from one medical center. No data recorded at that specific hospital was hereby included in the training and validation data. In an alternative setup, a balanced test set was randomly chosen from all patients. Hence, in this experiment training, validation and test set all included patients from each medical center available. This ensured a more diverse data composition.

### 3.3.5 Explainability

Means of creating a transparent and explainable algorithm were tried out and then meant to be compared considering several criteria. Given the complications that arose in attempting the implementation of several XAI tools, only the first part of this evaluation could be conducted in a comparative manner so far.

First, it is important to evaluate how practical or complicated the respective method is to be integrated into an existing algorithm. Hereby it is first considered how complex and time-consuming the required adaptations of the existing code are. Concurrently, it is beneficial to take into account if those changes and the implementation of each method tended to result in improved or decreased performance or showed other unanticipated side effects occur. This evaluation is also dependent on the type of adaptations necessary. Architectural changes of the network itself are for example considered as more complex and costly than adaptations relating to superficial code structure or arrangements that do not affect the network architecture.

The final objective of the XAI evaluation obviously concerns how well each approach is able to explain the algorithms decisions. Hereby the goals is to compare visual presentability and intuitive comprehensibility. For this evaluation the end user of such systems (e.g. doctors) needs to be kept in mind, while simultaneously considering which explanations are relevant for the designers and developers of the algorithm.

A more detailed experimental theme of XAI unfortunately exceeds the scope of this thesis. For now, the four techniques of ProtoPNet, SHAP, LRP and attention pooling

were inspected in some detail. In the experiments documented in Chapter 4, the latter was primarily applied and visualized in Section 4.1, while the attempt to implement the other three is discussed in the same section.

## 3.4 Statistical analysis

All of the experiments described above were statistically analysed and compared. This evaluation was mainly based on the statistical metrics of accuracy, F1-score, specificity and sensitivity, which will be defined in the following. For an overview over the foundational notions of *true positive*, *false positive*, *true negative* and *false negative*, a summary is found in the confusion matrix in Table 3. Hereby the notion *ground truth* indicates the label the algorithm learns from in a supervised learning setting, which for the issue at hand constitutes the official clinical diagnosis determined by a medical professional. *Classification* here means the prediction provided by the algorithm. Please note, that the official diagnosis can be flawed too and doesn't necessarily reflect the real ground truth (actually being sick with epilepsy or not). Hence, the matrix in 3 can also be interpreted considering the traditional diagnostic procedure and potential misdiagnoses.

Table 3: Confusion matrix explaining truth conditions using the example of this thesis topic.

| | |
|---|---|
| **True positive (TP)**<br>• Ground truth: epilepsy<br>• Prediction: epilepsy | **False positive (FP)** - Type I error<br>• Ground truth: healthy / no epilepsy<br>• Prediction: epilepsy |
| **False negative (FN)** - Type II error<br>• Ground truth: epilepsy<br>• Prediction: healthy / no epilepsy | **True negative (TN)**<br>• Ground truth: healthy / no epilepsy<br>• Prediction: healthy / no epilepsy |

Understanding and knowing the truth conditions is essential for calculating the following metrics for model evaluation.

*Accuracy* measures the number of predictions correctly classified by a model. This metric is quite basic and widely used for evaluating machine learning models. Unfortunately though, it misses out on valuable information in case of class-imbalanced data sets. For instance, a model might perform very poorly and simply classify every EEG recording as epileptic. If this model is evaluated using a validation set consisting of 20 epileptic but only 2 non-epileptic recordings, it will achieve an accuracy value of over 90%. Therefore,

usually other metrics are evaluated additionally to avoid misinterpretations.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

*Precision* is a metric for calculating what proportion of positive predictions were actually correct, namely how many true positives are predicted in relation to false positives.

$$Precision = \frac{True\ positives}{Total\ number\ of\ positive\ predictions} = \frac{TP}{TP + FP} \tag{5}$$

*Sensitivity*, or also referred to as *Recall*, measures how many actual positives were classified correctly.

$$Sensitivity = Recall = \frac{True\ positives}{Total\ number\ of\ actual\ positives} = \frac{TP}{TP + FN} \tag{6}$$

*Specificity*, as opposed to *Sensitivity*, calculates how many actual negatives were classified correctly.

$$Specificity = \frac{True\ negatives}{Total\ number\ actual\ negatives} = \frac{TN}{TN + FP} \tag{7}$$

Additionally, $F_1$ *score* was calculated as an alternative measure of *Accuracy*. It considers both *Precision* and *Sensitivity* by calculating the harmonic mean of those two measures. This fact enables the $F_1$ score to capture accuracy more meaningfully for imbalanced data sets with uneven class distribution.

$$F_1 score = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{8}$$

The above described metrics were calculated in the project at hand to measure the performance of the validation set during the training process, as well as for evaluating the independent test set at the end of the study.

# 4 Results

Table 4 shows a selection of conclusive experiments and their results. The listed results each resemble a single run of a specific experimental setup. Rather than repeating the exact same setup several times, the experiments in this thesis focused one evaluating as many different configurations as possible. From several hundred trials, seven informative example runs are displayed in Figure 4 for comparison. The most significant factors influencing the model performance are described in the following.

| # | **Architecture** | **Data** | **sec/w** | **Loss** | **LR** | **ValAcc** | **F1** | **Sens** | **Spec** | **TstAcc** |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BasicResNet | Raw, Dian. | 120 | MSE | 0.002 | 0.949 | 0.962 | 1.000 | 0.857 | 0.600 |
| 2 | BasicResNet | Pure, Dian. | 60 | Tukey | 0.002 | 0.917 | 0.929 | 0.867 | 1.000 | 0.627 |
| 3 | BasicResNet | Pure, rand. | 60 | Tukey | 0.02 | 0.923 | 0.950 | 1.000 | 0.714 | 0.820 |
| 4 | AttentionResNet | Pure, Dian. | 60 | Tukey | 0.0002 | 0.860 | 0.900 | 0.931 | 0.714 | 0.627 |
| 5 | AttentionResNet | Pure, Dian. | 120 | Tukey | 0.0001 | 0.907 | 0.933 | 0.966 | 0.786 | 0.627 |
| 6 | AttentionResNet | Pure, rand. | 180 | Tukey | 0.0001 | 0.881 | 0.912 | 0.929 | 0.786 | 0.610 |
| 7 | SimpleConvNet | Raw, Dian. | 60 | Tukey | 0.02 | 0.878 | 0.932 | 0.944 | 0.400 | 0.760 |

Table 4: Algorithm performance in various experimental setups. Column titles: Data = level of pre-processing of the input data with *Raw* being not pre-processed and *Pure* with automatic artifact reduction by PureEEG, *Dian.* denotes the use of the independent Dianalund test set, while *rand.* means training and test set were chosen at random from all data sources; sec/w= seconds per window of input EEG slices; LR = Learning rate; ValAcc = Best accuracy achieved on the validation set during training; Sens = Sensitivity, Spec = Specificity; TstAcc = Accuracy of evaluating the respective model with independent test data.

For explanation purposes, the model architectures experimented with are split into a simplified representation of three main types of networks. In reality small adaptations were made throughout the experimental process and networks in the same category are likely to be not entirely identical. Generally speaking the network categories are resembled by the graphics in Figure 13, 14 and 15 respectively. *SimpleConvNet* means a network without residual connections, but an architecture solely consisting of convolutional, pooling and dense layers (see Figure 13). Further types are referred to as *BasicResNet* and *AttentionResNet*, like it was differentiated in Section 3.2.2 and depicted in Figure 15 and

14.

The presented models were trained for up to 30 epochs. Each epoch took approximately 2-3 hours to train, depending on the number of training samples and the batch size. The number of steps per epoch corresponds to the number of samples divided by the batch size used in the respective experiment. During training, every 2000 steps the current model was evaluated using the validation set. According to these evaluation results early stopping was applied and the model that performed best on the validation set was saved. The validation accuracy values provided here correspond to the validation results of the saved model and the testing accuracy results from evaluating this best model afterwards using the test set. Figure 17 shows an example of how validation accuracy progresses over time during training. Figure 17 exhibits 18 epochs being evaluated over 39 hours.

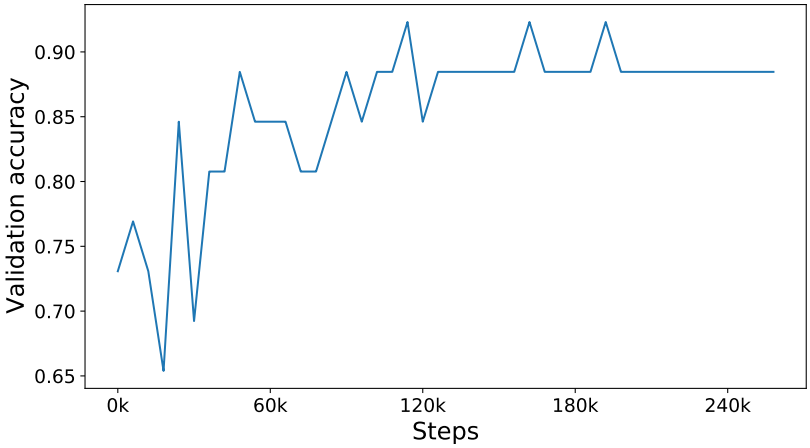


Figure 17: Example of accuracy development of the validation set over time during one training run (network described in #3 of Table 4).

Similar results were achievable for the BasicResNet and the rather shallow SimpleConvNet consisting of four convolutional layers after the SpatialCommon network. A notable insight when comparing the different architectures was, that suitable hyper-parameters were very individual and sensitive and needed careful adaptation when changing the architecture.

The learning rate in specific had to be adjusted carefully in order to maintain a high performance. With a learning rate between 0.02 and 0.0002 the network achieved the best results, but performance with identical learning rates varied greatly depending on the network architecture. In networks where an Attention Pooling layer was implemented, a significantly lower learning rate was crucial to make the model to learn at all. For an effective AttentionResNet a learning rate of at least 0.0002 or lower was necessary.

Tukey and MSE loss functions showed comparable results and the network performed significantly worse using the other functions. Thus, Tukey biweight function was imple-

mented in the final version of the algorithm, as it provided a more stable performance.

The effect of using slightly pre-processed data is not well visible in the plain results shown in Table 4. However, the networks turned to be more stable and consistent using 'pure' data and the highest performing networks were reached by training them with that type of data. Like already mentioned, the automatic tool PureEEG was conducted to produce those artifact reduced, clear EEG recordings. Regarding data choice it should be further noted that models generalized significantly better when the test set was randomly chosen from all patients (Table 4, #3). Opposed to this, using a test set consisting of recordings by an entirely independent data source (Danish Epilepsy Center) resulted in the model not being able to generalize to this data well at all.

Table 4 summarizes the networks with the presumably most important contribution factors, even though further variables played into the performance of the networks as well. Those factors were not included in the table to create a not too overloaded and well comprehensible overview. Anyhow, a quick overview will be given below.

Batch sizes were found to be effective at a minimum of 16 and a maximum of 32 patients per batch. Bigger batches tended to be less accurate and smaller batches were ineffective and used significantly larger amounts of computing power and time.

An increased number of input features improved the accuracy of the model, but more than 16 features were inefficient and immensely slowed down the training process and overloaded the GPU.

The optimal size to increase the feature depth in the SpatialCommon network and then per residual block was found at [1, 1.5, 3, 4.5, 6, 7.5, 9.0]. Deeper variations of [1, 1.5, 3, 4.5, 6, 9, 12.0] and [1, 1.5, 3, 6.0, 9.0, 12.0, 16.0] were experimented with as well, but did not show any positive effects.

## 4.1 Explainability

In the end, the implementation of several explainability tools had to be disregarded, due to compatibility issues with the utilized networks or the data type. Thus, this chapter first introduces the trials performed with the different XAI tools and their limitations. Finally it goes into detail and shows examples of the rather successful attention pooling method.

The application of the ProtoPNet turned out to be quite sophisticated. Given the described fundamental differences in problem formulation and data structure, applying this approach as proposed in the original paper was finally identified as unpromising. Exploring this issue further exceeded the scope of this thesis.

Further, also the implementation of the SHAP tool entailed an extensive period of trial and error and finally dashed against properly loading the pre-trained model. Investigation of this issue is beyond the timeframe and resources of this project, but will be subject to

future research.

The process of implementing LRP into the neural networks presented in this thesis already went further than the previously described methods. Unfortunately, also LRP did not yield any noteworthy results yet. The technique was simulated and inspected using the classical MNIST database, where LRP seemed like a great tool for explaining deep learning decisions in the image recognition field. The application to the clearly more complex EEG data used in this thesis did not succeed so far. For this to work, the existing networks have to be adapted entirely. Subsequently the hyperparameters again need to be optimized until the performance is considerably high for XAI tools to be applicable. This unfortunately exceeds the scope of this thesis. Further, it is not guaranteed to be successful with this data and research question, as LRP has been reported impractical with residual structure. Also, there is the chance that there are no other markers present in the non-ictal data besides spikes and there might be no novel insights from applying LRP. Therefore, ongoing research will investigate the application of LRP in automatic EEG analysis with different data and research questions first.

The presented AttentionResNet models finally made the exploration of the algorithms decisions and its reasoning possible. The plotting of the attention vector together with the input EEG recording slice, yields the results shown in Figures 18, 19 and 20. Hereby the attention vector shows in the form of a colored row below the corresponding EEG segment. This plot is set up in a way that areas of higher importance are highlighted in a brighter color. For example, in Figure 18 the green highlighting points to a clearly visible pattern in the EEG slice. Interestingly, it considers the rest of the recording piece as irrelevant even though two more distinct patterns can be visibly detected easily. In Figure 19 on the other hand, the light-green part marks a relatively long EEG period and hence does not point to detailed brain activity patterns. Figure 20 shows that in that particular EEG slice nothing was found to specifically play into the model's decision. This can mean that nothing in this sequence was informative or that everything was equally significant for the decision taken. As outlined in Section 3.2.3.3, this attention vector only highlights specific time points and does not specify the most significant electrode.

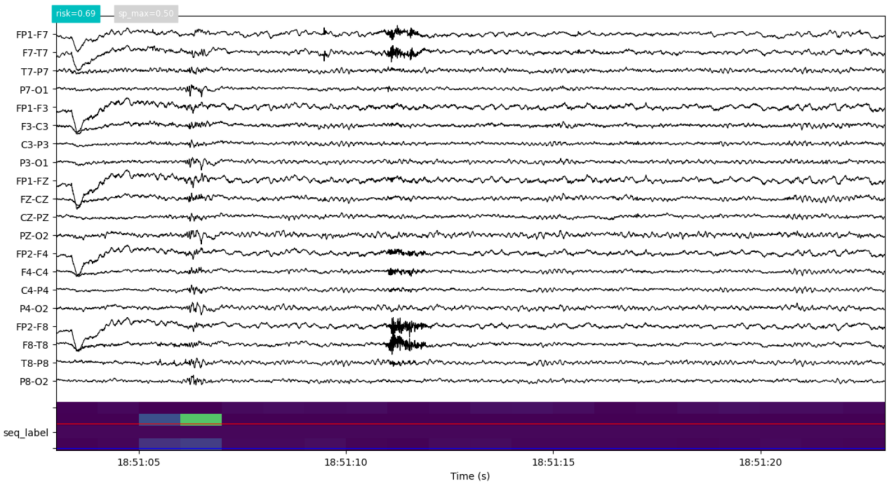


Figure 18: Attention vector pointing to abnormality in input EEG recording.

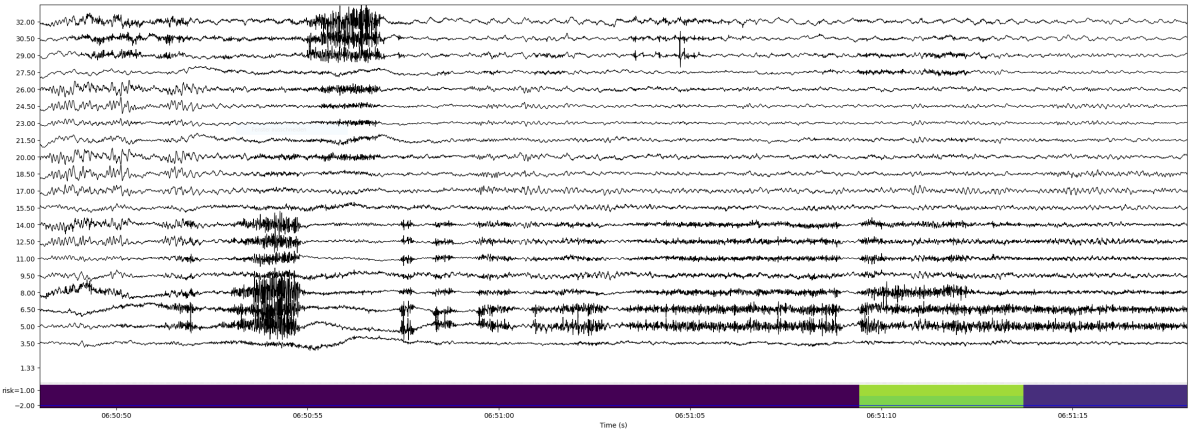


Figure 19: Attention vector not customized for size of input EEG slice.

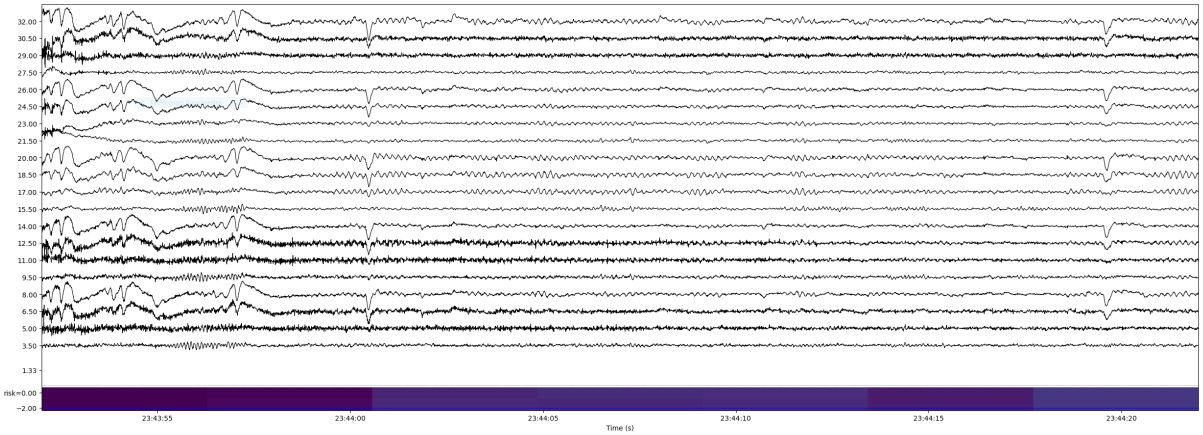


Figure 20: Attention vector on a regular, normal EEG slice.

# 5 Discussion and future work

The presented results of this AI modeling based thesis provide various conclusions from several interdisciplinary perspectives. It allows to reevaluate involved concepts and offers hints on future topics to be investigated. As the conducted experiments each take a lot of computational power and time to train, additional means of analysis unfortunately exceeded the scope and resources of this thesis project. The following paragraphs will also include several suggestions for future research in algorithmic epilepsy diagnosis.

Methodologically it can be concluded, that utilizing CNNs for evaluating raw EEG data is feasible, but the research questions and aims have to be thoughtfully chosen and clear. Otherwise, promising and reliable results cannot be produced. Within the course of the study presented here, it became evident that the initial aim of creating a direct end-to-end classifier for routine EEG data to assist epilepsy diagnosis potentially was too ambitious to easily create clinically valid results. Even though the achieved accuracy is informative and meaningful in scientific terms, it is not precise enough to be used in a clinical setting. When creating a medical tool for diagnostic use cases it is important to consider very high specificity as well as sensitivity scores as essential. High numbers of false positives can prevent the end user (medical professional/physician in charge) from gaining trust in the system and traditional diagnostic procedures would remain the preferred instrument. On the other hand, many false negatives lead to epileptics not being detected and sent home, which again entirely misses the point of the tool. Therefore, sensitivity and specificity play an essential role in analyzing the results presented in Chapter 4.

## 5.1 Impact of data choice

Besides remarks concerning the methodology and research questions of the presented study, also the choice of data sets for training and testing launches ideas for further discussion. First and foremost, the above presented results show the importance of testing neural network algorithms with an independent validation data set. Only because the model seems to exhibit a great performance, it is not implied that the right features are learned and the model is able to generalize well. The rather large gap between validation and testing accuracy that is present in many of the presented models (e.g. #1 in table 4) can be explained in various ways. One possibility could be that the algorithm was overfitting and hence could not generalize to an independent data set. On the other hand, it can also be suspected that the specific test data set could potentially be flawed. In the present case this option seems probable, given that the test data was recorded at a different hospital (Dianalund) than all recordings included in the training and validation sets. This theory is supported by further experiments, which were performed using randomly chosen recordings from all sources as a test set. Those experiments (e.g. #3 and #7 in Table 4)

achieved notably higher test performance compared to the ones tested on the Dianalund set. Thus, it can be concluded that a mixed data set, consisting of recordings provided by as many sources as possible, shows significantly better results than using an entirely independent test set by only one source. The network potentially is unable to classify the test data due to minor variances resulting from the individual recording procedure or device. Therefore, choosing the right data for training and testing that are equivalent but still entirely independent is essential here. To sum up, it can be stated that the presented results imply that ensuring variant and diverse data sets is crucial for successful, generalizable machine learning models.

## 5.2 Impact of network architectures and hyper-parameters

Regarding the application of CNNs and residual networks with EEG data, know-how was created that is applicable in further research. Generally, a more complex residual network with increased features is able to process more information. However, sometimes simpler, shallower networks seemingly create a better performance. However, it is considerably difficult to understand and explain what such networks are paying attention to and base their decisions on. Simply training a model without scrutinizing its results further can result in a network that is seemingly learning well, but not able to generalize well to a new data set. This effect is usually either due to overfitting on the training data or the network is not learning clinically valid information but some other high-level details (e.g. artifacts) found in the data that do not resemble brain activity. During the trials performed in this project, improving generalizable performance was accomplished through challenging default setups and experimenting with different structures and input data distributions. Hence, it can be learned from the presented results, that choosing an optimal model architecture should always go along with appropriate validation and testing. Thereby it is essential to utilize well chosen and diverse data sets, that represent all possible use cases of the algorithm as detailed as possible.

## 5.3 Implications for neuroscience

The initial aim of this thesis was to find ways in which a deep CNN will be able to detect biomarkers, which are not visually identifiable by human inspectors. This hypothesis was not confirmed by the performed experiments. Explainability mechanisms were hardly viable for this purpose and if so, they mainly pointed to spike biomarkers or seemingly random EEG slices being the reason for the algorithm's decisions. In that particular case a spike detection algorithm could be used instead and would presumably be more precise.

The presented results show, that the applied machine learning methods are not able to detect any novel hidden biomarkers in routine EEG recordings of epilepsy patients. This suggests that routine EEG without IEDs and seizures shows relatively normal brain

activity in epilepsy patients, that is remarkably similar to healthy brain activity. Hence, epilepsy might only manifest itself in ictal and interictal events, while the brain and neural activity otherwise resemble healthy behaviour in interictal periods. However, this thesis can only give evidence on the brain activity in short EEG slices of up to three minutes. Thus, the methods presented here fail to detect any biomarkers that might spread over larger time periods and do not show when only up to three minutes are inspected at once.

A more overarching biomarker detector might be achievable through approaching the question with a different type of analysis. Hereby one option could be to create an algorithm which is able to understand the data over a longer time frame, ideally creating a receptive field comprising the entire recording time. This could enable the network to detect activity patterns like changing oscillatory behaviour, recurrent patterns or coherence over time. Another attempt could be to give connectivity priority. Hereby, EEG recordings would be classified by analyzing if brain activity behaves synchronously and which brain areas are likely to be connected in different intensities depending on the diagnosis. Techniques like dynamic causal modeling (DCM) could be applied or the network could be trained on statistical connectivity measures utilizing e.g. phase locking value analysis.

If successful, this biomarker could then be combined with other networks focusing on specific features (like spike and HFO detection). Subsequently, a combined tool considering predefined promising features could potentially be developed. In a final stage this synergy could even be further combined with additional data that allows to make assumptions about epilepsy likelihood in each individual patient. Thus, information about seizure frequency, previous diseases, lifestyle and medication can be taken into account to create a holistic diagnosis assistant, in line with the precision medicine approach mentioned in the introduction.

## 5.4 Evaluation of XAI techniques

The presented results suggest that it is debatable if the current state-of-the-art of explainable AI tools provides the advertised solutions. While those tools are often claimed to be universal, many of them were designed only for very specific use cases. Implementation requires a lot of changes within the network architecture and a more challenging optimization process. In some cases the performance is impaired by excessive adaptations, which subsequently might motivate developers to disregard explainability to avoid the alternative of less accurate and potentially non-marketable models. Hence, from this perspective the target of explaining complex but efficient models is missed entirely.

Fortunately though, after lengthy trial and error periods attention pooling did provide relatively satisfying results. This approach can be reported as a quite convenient tool with the clear advantage of being integrated right into the network. After following some basic rules on choosing the structure and hyper-parameters, a network with included

attention is able to learn well and additionally provide some explanations. However, this type of explainability only yields well interpretable results when the size of the attention layer input equals the length (time in seconds) of the initial input it should explain. In the instant case this initial input corresponds to 60, 120 or 180 seconds of input EEG, which is convoluted through several residual blocks and should arrive at the attention pooling layer again with a size of 60, 120 or 180 respectively. This restriction requires intensive adjusting of the other parts of the network until the model yields acceptable results. Sometimes it is hardly possible to create results that are comparable to the performance of similar structures that do not satisfy this restriction. The consequences of this condition can be clearly seen in the visualizations in Chapter 4.1. While for example the model evidenced by #5 in table 4 showed great overall performance, it unfortunately does not offer a meaningful attention vector for visualization. This effect is due to the fact, that after several convolutional layers the size of the input data did not match with the original input length anymore and a considerably smaller attention vector is computed. This makes detailed interpretation like in Figure 18 impossible, as can be seen in Figure 19. On the other hand, well implemented attention pooling can create informative plots (Figure 18). Here the attention clearly points to salient patterns in the data.

Given the presented results and limitations, it also could be suspected that subtle epileptic patterns in EEG data might be too complex to be captured by attention pooling. Contrary to word sequences in NLP, IEDs are not identical, but only similar. Possibly, this similarity is not enough for the attention pooling to recognize the correspondence.

## 5.5 Implications for cognitive science

From the cognitive science perspective several conclusions can be drawn from this study that suggest improved strategies for future research in explainable AI. The topic of artificial intelligence and machine learning is intensely debated among cognitive scientists and philosophers. Hereby, ethical issue like AI alignment are made a priority and potential long-term dangers of AI are discussed. Also the topic of explainability is mainly tackled from this future-centered perspective. Despite AI alignment and artificial general intelligence being deeply interesting issues, they unfortunately seem to be almost entirely unrelated to current applications of state-of-the-art AI. As already mentioned in the introductory part of this thesis, cognitive science debates along the lines of explainability do not offer any applicable insights. However, I believe that its uniquely interdisciplinary orientation and knowledge base could provide highly relevant contributions. Neuroscience and psychology have already inspired many advances in AI like convolutional neural networks, residual networks or reinforcement learning. Likewise, insights into mechanisms like attention or understanding in the human brain could potentially be used to create biologically inspired explainable AI.

# 6 Conclusion

This project aimed to create a machine learning algorithm to assist epilepsy diagnosis from raw routine EEG data. This was ideally meant to be realized in an explainable and transparent manner. It can be concluded that there are many minor challenges to overcome until this main, wide-ranging goal can be reached. Open questions encompass more detailed data labelling, feature analysis and effective XAI tools for working with EEG data. Solving those issues requires consultation with medical experts for data labelling and the acquisition of greater amounts of data. Also more time and resources are necessary for exploring other potentially relevant features (e.g. connectivity, HFOs) and adapting XAI techniques to handle EEG data well. Tackling those solutions exceeded the scope of this thesis but offers promising opportunities for future research.

In general, this study shows that machine learning can be successfully applied to classification of epileptic routine EEG data with a validation accuracy during training of approximately 90% and a testing accuracy of 82%.

It can be further concluded, that the next aim in this specific research topic should be to create more stable results and increase the overall performance. Both could be done by introducing more data and by adapting or changing network architectures. Viewing larger EEG slices at once or focusing on different characteristics like connectivity could provide meaningful insights. However, also the possibility, that routine EEG recordings of epileptics might simply not provide many salient features besides the known IEDs (spikes, HFOs, etc.), should be considered likely. In this case, focusing further on creating more reliable algorithms for automatically detecting specific IEDs seems to offer the most reasonable approach.

Despite this being a request entailing much effort, more detailed labels could provide great opportunities for further investigation. Instead of labeling each recording as one option of a binary decision, abnormal activity could be annotated right when it occurs within the recordings. With such labels, the CNN could specifically learn what an abnormal or epileptic pattern looks like. In the case of overall patient labels on the other hand, it could be beneficial to train a regression of epilepsy risk or ictogenesis. The latter could be approached in future research by cutting long-term VEM recordings before seizures versus where the EEG just continues normally and labeling them accordingly. Then a seizure prediction algorithm could be attempted.

On a more general account, it can be reasoned that the discussed area of research could profit more from its high interdisciplinarity. A close collaboration on equal terms combining the fields of computer science, neuroscience, psychology and philosophy could provide great insights and know-how.

# Appendices

## A Code snippets

Listing 1: Python code showing the residual blocks the residual network is built of. Block type (Conv. or Identity) and filter/kernel/stride sizes are defined when calling this function.

```
import tensorflow as tf

def Resnetv2_block(x, filters, growfilters, kernel_size=3, stride=1,
    conv_shortcut=False, name=None):
    k = tf.keras.layers

    if conv_shortcut is True:
        shortcut = ConvSepBatch(x, growfilters, kernel_size=1, stride=
            stride, name=name+"_sc")
    else:
        shortcut = x

    x = k.Conv2D(filters=filters, kernel_size=(1,1), strides=(1,1), padding
        ='same', activation=None, name=name + "_c1")(x)
    x = k.BatchNormalization(name=name + "_bn")(x)
    x = k.Activation('relu', name=name + "_act1")(x)

    x = k.Conv2D(filters=filters, kernel_size=(kernel_size,1), strides=(
        stride,1), padding='same', activation=None, name=name + "_c2")(x)
    x = k.BatchNormalization(name=name + "_bn")(x)
    x = k.Activation('relu', name=name + "_act2")(x)

    x = k.Conv2D(filters=filters, kernel_size=(kernel_size,1), strides=(
        stride,1), padding='same', activation=None, name=name + "_c3")(x)
    x = k.BatchNormalization(name=name + "_bn")(x)
    x = k.Activation('relu', name=name + "_out")(x + shortcut)

    return x
```

Listing 2: Python code the SpatialCommon network used to create spatial features and reduce the input data to 2D representation.

```
    def SpatialCommon(self, input):

        k = tf.keras.layers
        input3d = tf.expand_dims(input, -1)
```

```
        first_filter = self.config['cnn_filters_bio']

        x = k.Conv3D(filters=int(first_filter*self.config['common_phi'][0])
            , kernel_size=(7, 3, 1), strides=(1, 1, 1), padding='same',
            activation=None, name='com_c1')(input3d)
        chfeat = self.channelFeatureDyn(input3d, tf.shape(x))
        x = k.concatenate([x, chfeat], axis=4, name='CONCAT_CHFEAT')
        x = k.MaxPool3D(pool_size=(4, 1, 1), strides=(4, 1, 1), padding='
            valid', name='com_m1')(x)

        x = k.Conv3D(filters=int(first_filter*self.config['common_phi'][1])
            , kernel_size=(3, 3, 1), strides=(1, 1, 1), padding='same',
            activation=tf.nn.relu, name='com_c2')(x)
        x = k.MaxPool3D(pool_size=(2, 3, 1), strides=(2, 3, 1), padding='
            same', name='com_m2')(x)

        x = tf.squeeze(x, 2)  # slices evaluated by convnet resulting in
            slice features, remove dimension
        return x
```

Listing 3: Configuration of network architecture setup and hyper-paramenters of the final BasicResNet model

```
config['backb_level'] = 3
config['cnn_filters_spat'] = 16
config['cnn_filters_chan'] = 2
config['fc_layer_size'] = 512
config['secondsPerWindow'] = 60*1
config['fs_for_learning'] = 128
config['cnn_filters_bio'] = 16
config['algtest'] = 3
config['common_phi'] = [1.0, 1.5]
config['resnet_blocks'] = [3,3,3,3,1]
config['resnet_depthinc'] = [3.0, 4.5, 6.0, 7.5, 9.0]
config['num_layers'] = 1
config['num_heads'] = 2
config['d_model'] = 64
config['learning_rate'] = 0.02
config['lr_decay_rate'] = 0.98
config['lr_decay_steps'] = 700
config['warmup_steps'] = 200
config['batch_size'] = 16
config['NchSlice'] = 3
config['epochs'] = 30
```

Listing 4: Code for implementing Tukey biweight loss function.

```
def TukeysLoss(self, mindist):
```

```
    c = 1

    slope = 0.01
    r = tf.abs(mindist)
    pforlarge = tf.ones_like(r) * (c ** 2) / 6 + r * slope + slope
    p = pforlarge * (1 - (1 - (r / c) ** 2) ** 3) + r * slope

    comparision = tf.less_equal(r, c)
    loss = tf.where(comparision, p, pforlarge)

    return loss
```

Listing 5: Example output of residual network model evaluation on independent test set.

```
runtime: 1208 seconds EEG in 6.59 sec : 7.9 min for 24h EEG
Epilepsy risk ########################## :1.000 for Ground truth: 1
runtime: 1216 seconds EEG in 6.02 sec : 7.1 min for 24h EEG
Epilepsy risk ########################## :1.000 for Ground truth: 1
runtime: 1225 seconds EEG in 7.23 sec : 8.5 min for 24h EEG
Epilepsy risk ######                    :0.339 for Ground truth: 0
runtime: 739 seconds EEG in 6.82 sec : 13.3 min for 24h EEG
Epilepsy risk #                         :0.075 for Ground truth: 0
runtime: 1205 seconds EEG in 6.41 sec : 7.7 min for 24h EEG
Epilepsy risk ########################## :1.000 for Ground truth: 1
runtime: 1217 seconds EEG in 5.02 sec : 5.9 min for 24h EEG
Epilepsy risk ########################## :1.000 for Ground truth: 0
runtime: 1217 seconds EEG in 5.55 sec : 6.6 min for 24h EEG
Epilepsy risk ########################  :0.976 for Ground truth: 1
runtime: 1212 seconds EEG in 5.80 sec : 6.9 min for 24h EEG
Epilepsy risk ################          :0.779 for Ground truth: 1
runtime: 1219 seconds EEG in 5.02 sec : 5.9 min for 24h EEG
Epilepsy risk                           :0.000 for Ground truth: 0
runtime: 1232 seconds EEG in 6.12 sec : 7.2 min for 24h EEG
Epilepsy risk ########################## :1.000 for Ground truth: 1
```

# References


Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.*

Abbasi, B., & Goldenholz, D. M. (2019). Machine learning applications in epilepsy. *Epilepsia*, *60*(10), 2037–2047.

Acharya, J. N., Hani, A. J., Cheek, J., Thirumala, P., & Tsuchida, T. N. (2016). American Clinical Neurophysiology Society Guideline 2: Guidelines for Standard Electrode Position Nomenclature. *The Neurodiagnostic Journal*, *56*(4), 245–252.

Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., ... Friedman, P. A. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, *394*(10201), 861–867.

Bagheri, E., Jin, J., Dauwels, J., Cash, S., & Westover, M. B. (2019). A fast machine learning approach to facilitate the detection of interictal epileptiform discharges in the scalp electroencephalogram. *Journal of Neuroscience Methods*, *326*, 108362.

Baumgartner, C. (2001). *Handbuch der epilepsien: Klinik, diagnostik, therapie und psychosoziale aspekte.* Springer Vienna.

Baumgartner, C., & Pirker, S. (2019). Video-EEG. In *Handbook of Clinical Neurology* (Vol. 160, pp. 171–183). Elsevier.

Benbadis, S. R., LaFrance, W. C., Papandonatos, G. D., Korabathina, K., Lin, K., Kraemer, H. C., & For the NES Treatment Workshop. (2009). Interrater reliability of EEG-video monitoring. *Neurology*, *73*(11), 843–846.

Boran, E., Sarnthein, J., Krayenbühl, N., Ramantani, G., & Fedele, T. (2019). High-frequency oscillations in scalp EEG mirror seizure frequency in pediatric focal epilepsy. *Scientific Reports*, *9*(1), 16560.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., ... Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, *25*(8), 1301–1309.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission.

In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). Sydney, NSW, Australia: Association for Computing Machinery.

Centeno, M., & Carmichael, D. (2014). Network connectivity in epilepsy: Resting state fmri and eeg-fmri contributions. *Frontiers in Neurology*, *5*, 93.

Chen, C., Li, O., Barnett, A., Su, J., & Rudin, C. (2018). This looks like that: deep learning for interpretable image recognition. *CoRR*, *abs/1806.10574*.

Chen, D. K., Sharma, E., & LaFrance, W. C. (2017). Psychogenic Non-Epileptic Seizures. *Current Neurology and Neuroscience Reports*, *17*(9), 71.

de Curtis, M., Jefferys, J. G. R., & Avoli, M. (2012). Interictal Epileptiform Discharges in Partial Epilepsy: Complex Neurobiological Mechanisms Based on Experimental and Clinical Evidence. In J. L. Noebels, M. Avoli, M. A. Rogawski, R. W. Olsen, & A. V. Delgado-Escueta (Eds.), *Jasper's Basic Mechanisms of the Epilepsies* (4th ed.). Bethesda (MD): National Center for Biotechnology Information (US).

Devinsky, O., Gazzola, D., & LaFrance, W. C., Jr. (2011). Differentiating between nonepileptic and epileptic seizures. *Nature Reviews Neurology*, *7*(4), 210–.

Doss, R. C., & LaFrance, W. C. (2016). Psychogenic non-epileptic seizures. *Epileptic Disorders*, *18*(4), 337–343.

Engel, J., Bragin, A., & Staba, R. (2018). Nonictal EEG biomarkers for diagnosis and treatment. *Epilepsia Open*, *3*(Suppl Suppl 2), 120–126.

Fedele, T., Schönenberger, C., Curio, G., Serra, C., Krayenbühl, N., & Sarnthein, J. (2017). Intraoperative subdural low-noise EEG recording of the high frequency oscillation in the somatosensory evoked potential. *Clinical Neurophysiology*, *128*(10), 1851–1857.

Fellous, J.-M., Sapiro, G., Rossi, A., Mayberg, H., & Ferrante, M. (2019). Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation. *Frontiers in Neuroscience*, *13*, 1346.

Fisher, R. S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J. H., Elger, C. E., ... Wiebe, S. (2014). ILAE Official Report: A practical clinical definition of epilepsy. *Epilepsia*, *55*(4), 475–482.

Fisher, R. S., Boas, W. v. E., Blume, W., Elger, C., Genton, P., Lee, P., & Engel, J. (2005). Epileptic Seizures and Epilepsy: Definitions Proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). *Epilepsia*, *46*(4), 470–472.

Fisher, R. S., Cross, J. H., French, J. A., Higurashi, N., Hirsch, E., Jansen, F. E., ... Zuberi, S. M. (2017). Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. *Epilepsia*, *58*(4), 522–530.

Fodor, J. A. (1975). *The language of thought* (Repr.. ed.). Hassocks: Harvester Press.

Frauscher, B., Bartolomei, F., Kobayashi, K., Cimbalnik, J., van 't Klooster, M. A., Rampp, S., ... Gotman, J. (2017). High-frequency oscillations: The state of clinical research. *Epilepsia*, *58*(8), 1316–1329.

Frauscher, B., von Ellenrieder, N., Zelmann, R., Rogers, C., Nguyen, D. K., Kahane, P., ... Gotman, J. (2018). High-Frequency Oscillations in the Normal Human Brain: HFOs in the Normal Human Brain. *Annals of Neurology*, *84*(3), 374–385.

Fu, X., Wang, Y., Ge, M., Wang, D., Gao, R., Wang, L., ... Liu, H. (2018). Negative effects of interictal spikes on theta rhythm in human temporal lobe epilepsy. *Epilepsy & Behavior: E&B*, *87*, 207–212.

Fürbass, F., Kural, M. A., Gritsch, G., Hartmann, M., Kluge, T., & Beniczky, S. (2020). An artificial intelligence-based EEG algorithm for detection of epileptiform EEG discharges: Validation against the diagnostic gold standard. *Clinical Neurophysiology*, *131*(6), 1174–1179.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:1806.00069 [cs, stat]*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grant, A. C., Abdel-Baki, S. G., Weedon, J., Arnedo, V., Chari, G., Koziorynska, E., ... Omurtag, A. (2014). EEG interpretation reliability and interpreter confidence: A large single-center study. *Epilepsy & Behavior*, *32*, 102–107.

Gschwandtner, L. (2020). Deep learning for estimation of functional brain maturation from EEG of premature neonates. *42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society*, *Submitted*.

Halford, J. J. (2009). Computerized epileptiform transient detection in the scalp electroencephalogram: Obstacles to progress and the example of computerized ECG interpretation. *Clinical Neurophysiology*, *120*(11), 1909–1915.

Hartmann, M. M., Schindler, K., Gebbink, T. A., Gritsch, G., & Kluge, T. (2014). PureEEG: automatic EEG artifact removal for epilepsy monitoring. *Neurophysiologie Clinique = Clinical Neurophysiology*, *44*(5), 479–490.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

Hopp, J. L. (2019). Nonepileptic Episodic Events:. *CONTINUUM: Lifelong Learning in Neurology*, *25*(2), 492–507.

Höller, P., Trinka, E., & Höller, Y. (2018). High-Frequency Oscillations in the Scalp Electroencephalogram: Mission Impossible without Computational Intelligence. *Computational Intelligence and Neuroscience*, *2018*.

Höller, Y., Trinka, E., Kalss, G., Schiepek, G., & Michaelis, R. (2019). Correlation of EEG spectra, connectivity, and information theoretical biomarkers with psychological states in the epilepsy monitoring unit - A pilot study. *Epilepsy & Behavior: E&B*, *99*, 106485.

Jing, J., Herlopian, A., Karakis, I., Ng, M., Halford, J. J., Lam, A., ... Westover, M. B. (2020). Interrater Reliability of Experts in Identifying Interictal Epileptiform Discharges in Electroencephalograms. *JAMA Neurology*, *77*(1), 49.

Jing, J., Sun, H., Kim, J. A., Herlopian, A., Karakis, I., Ng, M., ... Westover, M. B. (2020). Development of Expert-Level Automated Detection of Epileptiform Discharges During Electroencephalogram Interpretation. *JAMA Neurology*, *77*(1), 103.

Jing, L., & Tian, Y. (2019). Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *arXiv:1902.06162 [cs]*.

Kearney, H., Byrne, S., Cavalleri, G. L., & Delanty, N. (2019). Tackling Epilepsy With High-definition Precision Medicine: A Review. *JAMA Neurology*, *76*(9), 1109-1116.

Knezevic, C. E., & Marzinke, M. A. (2018). Clinical Use and Monitoring of Antiepileptic Drugs. *The Journal of Applied Laboratory Medicine*, *3*(1), 115-127.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R., & Samek, W. (2016). The lrp toolbox for artificial neural networks. *Journal of Machine Learning Research*, *17*(114), 1-5.

Lu, D., & Triesch, J. (2019). *Residual Deep Convolutional Neural Network for EEG Signal Classification in Epilepsy.*

Luders, H. O. (2008). *Textbook of Epilepsy Surgery.* CRC Press.

Lugo-Reyes, S. O., Maldonado-Colín, G., & Murata, C. (2014). [Artificial intelligence to assist clinical diagnosis in medicine]. *Revista Alergia Mexico (Tecamachalco, Puebla, Mexico: 1993)*, *61*(2), 110–120.

Lundberg, S. M., & Lee, S.-I. (n.d.). A Unified Approach to Interpreting Model Predictions. , 10.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, *20*(5), e262–e273.

Panayiotopoulos, C. P. (2010). *A Clinical Guide to Epileptic Syndromes and their Treatment*. Springer Science & Business Media.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, ca, usa, august 13-17, 2016* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). San Francisco California USA: ACM.

Rojas, I., Joya, G., & Catala, A. (2019). *Advances in computational intelligence: 15th international work-conference on artificial neural networks, iwann 2019, gran canaria, spain, june 12-14, 2019, proceedings, part ii*. Springer International Publishing.

Russell, S. J. (2014). *Artificial intelligence : a modern approach* (3. ed., Pearson new international ed.. ed.). Harlow: Pearson Education.

Scheffer, I. E., Berkovic, S., Capovilla, G., Connolly, M. B., French, J., Guilhoto, L., ... Zuberi, S. M. (2017). ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. *Epilepsia*, *58*(4), 512–521.

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*, *9*(1), 12495.

Sinha, S. R., Sullivan, L. R., Sabau, D., Orta, D. S. J., Dombrowski, K. E., Halford, J. J., ... Stecker, M. M. (2016). American Clinical Neurophysiology Society Guideline 1: Minimum Technical Requirements for Performing Clinical Electroencephalography. *The Neurodiagnostic Journal*, *56*(4), 235–244.

Staba, R. J., Stead, M., & Worrell, G. A. (2014). Electrophysiological Biomarkers of Epilepsy. *Neurotherapeutics*, *11*(2), 334–346.

Sturm, I., Bach, S., Samek, W., & Müller, K.-R. (2016). Interpretable Deep Neural Networks for Single-Trial EEG Classification. *arXiv:1604.08201 [cs, stat]*.

Tjepkema-Cloostermans, M. C., de Carvalho, R. C., & van Putten, M. J. (2018). Deep learning for detection of focal epileptiform discharges from scalp EEG recordings. *Clinical Neurophysiology*, *129*(10), 2191–2196.

Trinh, T. H., Luong, M., & Le, Q. V. (2019). Selfie: Self-supervised pretraining for image embedding. *CoRR*, *abs/1906.02940*.

van Leeuwen, K., Sun, H., Tabaeizadeh, M., Struck, A., van Putten, M., & Westover, M. (2019). Detecting abnormal electroencephalograms using deep convolutional networks. *Clinical Neurophysiology*, *130*(1), 77–84.

van Mierlo, P., Papadopoulou, M., Carrette, E., Boon, P., Vandenberghe, S., Vonck, K., & Marinazzo, D. (2014). Functional brain connectivity from EEG in epilepsy: seizure prediction and epileptogenic focus localization. *Progress in Neurobiology*, *121*, 19–35.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*.

Weaver, D. F. (2004). "Organic" pseudoseizures as an unrecognized side-effect of anticonvulsant therapy. *Seizure*, *13*(7), 467–469.

Westberg, M., Zelvelder, A., & Najjar, A. (2019). A historical perspective on cognitive science and its influence on xai research. In (p. 205-219).

World Health Organization. (2019). *Epilepsy.* `https://www.who.int/news-room/fact-sheets/detail/epilepsy`.

Yin, C., Zhang, X., Xiang, J., Chen, Z., Li, X., Wu, S., ... Wang, Y. (2019). Altered effective connectivity network in patients with insular epilepsy: A high-frequency oscillations magnetoencephalography study. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *131*(2), 377–384.

Yin, W., & Schütze, H. (2018). Attentive Convolution: Equipping CNNs with RNN-style Attention Mechanisms. *Transactions of the Association for Computational Linguistics*, *6*, 687–702.

Zhuo Ding, J., Mallick, R., Carpentier, J., McBain, K., Gaspard, N., Brandon Westover, M., & Fantaneanu, T. A. (2019). Resident training and interrater agreements using the ACNS critical care EEG terminology. *Seizure*, *66*, 76–80.

Zibrandtsen, I. C., Weisdorf, S., Ballegaard, M., Beniczky, S., & Kjaer, T. W. (2019). Postictal EEG changes following focal seizures: Interrater agreement and comparison to frequency analysis. *Clinical Neurophysiology*, *130*(6), 879–885.

Zschocke, S., & Kursawe, H. K. (2012). *Klinische elektroenzephalographie* (3., aktual. und erw. Aufl.. ed.). Berlin [u.a.]: Springer.

Zweiphenning, W. J. E. M., van Diessen, E., Aarnoutse, E. J., Leijten, F. S. S., van Rijen, P. C., Braun, K. P. J., & Zijlmans, M. (2020). The resolution revolution: Comparing spikes and high frequency oscillations in high-density and standard intra-operative electrocorticography of the same patient. *Clinical Neurophysiology*, *131*(5), 1040–1043.