FACULTY OF ENGINEERING OF THE UNIVERSITY OF PORTO

Deep Learning for EEG Analysis in Epilepsy

Catarina da Silva Lourenço

INTEGRATED MASTERS IN BIOENGINEERING

Supervisor: Prof. Dr. Michel van Putten

Supervisor: Prof. Dr. Luís Teixeira

June 23rd 2019

Deep Learning for EEG Analysis in Epilepsy

Catarina da Silva Lourenço

INTEGRATED MASTERS IN BIOENGINEERING

June 23rd 2019

Abstract

Epilepsy is a brain disease that entails a predisposition to generate seizures. Electroencephalography (EEG) is currently the gold standard for diagnosing epilepsy. Since the availability of ictal EEGs is scarce, diagnosis is often done by experts based on visual analysis of interictal periods. These are characterized by the occurrence of interictal epileptiform discharges (IEDs). While this is still the gold standard, it entails several disadvantages that motivate the need for developing algorithms that automate the process, reducing subjectivity and the time spent by experts on diagnosis. Given that deep learning is unbiased towards the features currently used in visual inspection and is able to learn from raw data, it can be an alternative to visual inspection and traditional machine learning methods for EEG analysis.

We trained four convolutional models (VGG, ResNet and two custom-made models) using 2-second IED epochs from patients with focal and generalized Epilepsy (39 and 40 patients, respectively, 1977 epochs total), as well as normal EEGs from controls (110770 epochs from 53 controls). Five-fold cross-validation was performed on the training set and testing was done on an independent set (734 epochs with IEDs from 10 patients, 23040 normal epochs from 14 controls). We calculated the average ROC curves and corresponding areas under the curves (AUC). Sensitivity, specificity, true positive and false positive rates were assessed at several thresholds. Filter visualization, input maximization and occlusion were applied to gather information about the behavior of the models.

The VGG model led to the best results in this task. It yielded an AUC of 0.96 (Confidence Interval at 95% (CI)=0.95-0.97) on the test set. At a threshold of 0.5, every EEG of the normal class was classified with a specificity value over 95%, with four files reaching 100% specificity. It led to an average sensitivity of 93% and specificity of 91% on the files containing IEDs. The intersection between the sensitivity and specificity values on the test set was 93%, with 122.41 (CI=27.63-217.20) false positives and 32.31 (CI=15.15-49.46) true positives per hour. At 99% specificity, one sample was misclassified per 2 minutes of EEG. Filter visualization showed differences between filters of lower and higher level layers of the VGG, with higher level layers showing patches corresponding to IED detection. Occlusion showed that IED shapes were being clearly identified by the network in true positive cases, showing that the network detected the correct patterns and not spurious features, making the model more empirically reliable.

We prove the potential of deep learning techniques in IED detection. This work is innovative in its use of deep networks described in the literature, which are widely used in other areas but not yet in the scope of IED detection. Furthermore, it showcases the usefulness of visualization techniques in illustrating some of the processes behind classification, which is also new in the paradigm of deep learning in EEG analysis.

Resumo

A Epilepsia é uma doença neurológica que envolve uma predisposição para a ocorrência de ataques epiléticos. O eletroencefalograma (EEG) é atualmente a técnica usada para o diagnóstico da epilepsia. Dado que a aquisição do EEG ictal dos pacientes não é comum, o diagnóstico é muitas vezes feito com base na análise visual de períodos interictais, caracterizados pela ocorrência de descargas epileptiformes interictais (IEDs). Embora a análise visual seja o estado da arte, esta técnica tem desvantagens que motivam o desenvolvimento de algoritmos que automatizem o processo. Os métodos de *deep learning* não são enviesados pelas características a que os clínicos recorrem para a classificação dos sinais de EEG e são capazes de aprender diretamente a partir de dados. Assim, estes métodos podem ser uma alternativa à inspeção visual e aos métodos tradicionais de *machine learning* para análise de EEGs.

No âmbito deste projeto foram treinadas quatro redes neuronais convolucionais (VGG, ResNet e dois modelos desenvolvidos pelo grupo de investigação). Para o treino dos modelos, foram usadas 1977 amostras de IEDs com a duração de 2 segundos (provenientes de 39 pacientes com Epilepsia focal e 40 pacientes com Epilepsia generalizada), bem como 110770 amostras de EEGs normais (de 53 controlos). Foi aplicada *cross-validation* nos dados de treino e o modelo foi validado em dados independentes (734 IEDs de 10 pacientes e 23040 amostras de 14 controlos). Calculou-se a curva ROC média das cinco iterações de *cross-validation*, bem como o valor da área abaixo desta curva (AUC). Determinaram-se, para vários limiares, os valores da sensibilidade, especificidade e as taxas de verdadeiros e falsos positivos por hora. Foram também aplicadas técnicas de visualização dos modelos (*filter visualization*, *input maximization* e *occlusion*).

A rede VGG obteve os melhores resultados, com uma AUC de 0.96 (intervalo de confiança a 95% (CI)=0.95-0.97) nos dados de teste. Com um limiar de classificação de 0.5, todos os EEGs de teste da classe normal foram classificados com uma especificidade superior a 95%, sendo que quatro foram classificados com 100% de especificidade. Obtiveram-se uma sensibilidade e especificidades médias de 93% e 91%, respetivamente, nos ficheiros com IEDs. A interseção dos valores de sensibilidade e especificidade foi 93%, com 122.41 (CI=27.63-217.20) falsos positivos por hora e 32.31 (CI=15.15-49.46) verdadeiros positivos por hora. Apenas uma amostra foi incorretamente classificada em cada 2 minutos de EEG, a uma especificidade de 99%.

A técnica de *filter visualization* mostrou que há diferenças visíveis entre filtros de diferentes camadas da rede VGG, sendo que os filtros das camadas de níveis mais elevados apresentam zonas de atividade correspondentes à deteção de IEDs. A técnica de *occlusion* mostrou que as formas dos IEDs foram corretamente identificadas pelas redes, tornando o modelo mais empiricamente fiável.

Com este trabalho, foi possível provar o potencial das técnicas de *deep learning* para a deteção de IEDs, inovando pela aplicação de redes neuronais descritas na literatura e vastamente usadas noutras áreas, mas não neste âmbito. Adicionalmente, demonstrou-se a utilidade das técnicas de visualização de redes neuronais, constituindo também uma novidade no paradigma da deteção de IEDs.

Acknowledgments

I have received much support and assistance during the development of this thesis, for which I am very grateful.

I would first like to thank Michel van Putten and Marleen Tjepkema-Cloostermans for their support, availability, assistance and the trust they gave me and my work in the CNPH group since day one. Their expertise and knowledge of the clinical and technical domains has been invaluable in the development of this project and their guidance was crucial in this dissertation.

I would like to thank Luís Teixeira for his support, ideas and valuable feedback throughout the development of this thesis, as well as for the wonderful introduction to Deep Learning, which couldn't have come at a better time. I would also like to thank Christin Seifert and Meike Nauta for their feedback and ideas. I hope our collaboration grows stronger over these upcoming years.

I want to thank everyone in the CNPH group for making me feel welcome and creating the best work environment one could ask for. Julia, Marloes and Joliene (albeit not currently in CNPH), I am truly grateful for your friendship, support and help with all sorts of (usually very random) matters. I would also like to thank Tanja, for all the help and support since my very first day at CNPH, as well as Annika, Sophie, Monica and Sjoukje.

I am incredibly thankful to my parents for the principles they conveyed during my formative years, which undoubtedly shaped me into who I am, for putting up with me and for their support in what concerns my work and life in general. I want to thank my whole family, whose unwavering love and support are always on my mind and André, for the crucial encouragement and kindness, for always being present and for not letting me stress out too much.

Thank you all.

Catarina Lourenço

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Context

Epilepsy is the fourth most prevalent neurological disorder in the world. It is a disease of the brain that entails a predisposition to generate seizures, encompassing a plethora of syndromes and clinical phenomenology, some similar to other diseases [1–3]. Thus, distinguishing a non-epileptic paroxysmal event from a seizure is clinically difficult, and the rate of misdiagnosis for epilepsy is reported to be up to 30% [4, 5]. This may result in an increased risk of recurrent seizures due to lack of adequate treatment or prescription of potentially harmful medication to patients with other disorders [6, 7]. Accurate and timely diagnosis, therefore, is clinically highly relevant. Ideally, this should be done in an automated way to reduce subjectivity and the time that experts spend on this type of diagnosis.

EEG is one of the most useful techniques for diagnostics in epilepsy and classification of epilepsy syndromes [1, 8]. Ictal EEG, i.e. the EEG measured during a seizure, is the only method that nearly always unequivocally distinguishes an epileptic seizure from a non-epileptic one, allowing certain diagnosis of the disease. However, the likelihood of acquiring an ictal EEG is low due to the unpredictability of seizure occurrence [9, 10]. In many patients, the interictal EEG shows Interictal Epileptiform Discharges (IEDs): transient patterns that indicate an increased likelihood of seizures. These IEDs help to differentiate epilepsy from other conditions [11, 12]. Assessment of their presence is done by visual analysis, which has been the gold standard in the clinic for almost over a century [13]. Yet, the learning curve is long, review times are significant, visual assessment is subjective and inter and intra-individual variability ranges from 5 to 25% [14, 15]. Despite these limitations, visual assessment of the EEG still outperforms current computer algorithms in detecting IEDs.

There is a clear need for the development of algorithms that are able to match or outperform visual assessment of EEG data in what concerns the detection of IEDs for the timely and assertive diagnosis of epilepsy. These should use as input the raw signal, being unbiased regarding the features that are usually extracted in visual analysis, since this process may fail to capture important information due to the richness of the EEG signal. Thus, Deep Learning methods appear to be

a viable potential solution for this problem. While the application of these techniques to fields such as image analysis is already well established, the use of Deep Learning in health, and in IED detection in particular, is now starting to grow, making this topic of research even more relevant.

1.2 Structure of the Dissertation

This document is divided into 9 chapters. This first chapter presents the topic of the dissertation, as well as the motivation for researching this subject.

Chapter 2 covers the history of the EEG, the acquisition of this type of signal as well as the signal itself. EEG analysis and its clinical applications are also discussed.

Chapter 3 focuses on Epilepsy, including a historical overview, as well as remarks regarding the aetiology of the disease and its manifestation through seizures. The role of the EEG signal in the diagnosis process and the available treatments for the disease are also covered in this chapter.

Chapter 4 provides a broad overview of Deep Learning models and of the history of Deep Learning itself. It covers the different types of learning used for Deep Learning tasks, as well as how to assess performance and gain insight into the training process. Finally, the clinical applications of Deep Learning are discussed.

Chapter 5 describes the State of the Art of Machine Learning methods in Epilepsy, covering seizure detection and prediction, treatment optimization and IED detection.

Chapter 6 covers the methods that were implemented and used in the dissertation and Chapter 7 describes the results obtained with said methods. Chapter 8 concerns the discussion of the results presented in Chapter 7.

Finally, Chapter 9 presents the main conclusions that can be drawn from this dissertation and describes several ideas and steps for future work that will be carried out in the upcoming months.

Chapter 2

Electroencephalography

Electroencephalography (EEG) consists in the electrophysiological recording of electrical activity produced by the brain. The neural activity detectable in the EEG is the summation of excitatory and inhibitory postsynaptic potentials generated by groups of synchronously firing cortical pyramidal cells oriented perpendicularly to the brain's surface [16–18]. By plotting the recorded voltages against time, it is possible to obtain a display of large scale neural dynamics [19, 20].

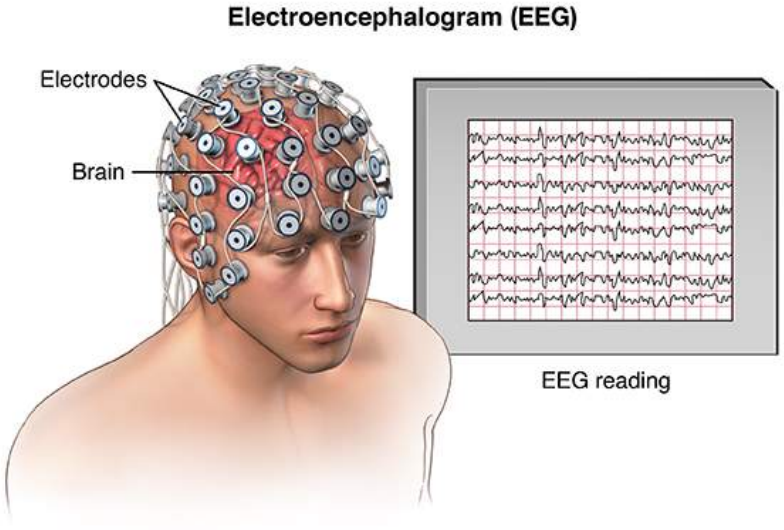


Figure 2.1: Cartoon showing electrode positioning on the scalp and the electrical waves recorded in the EEG [21].

2.1 Historical Perspective

The first neurophysiological recordings were performed by Richard Caton in 1875 on the exposed scalp of rabbits and monkeys [22, 23]. However, the first recording of a human EEG was only made in 1924 by Hans Berger, using non-polarizable clay cylinder electrodes and a string galvanometer [24, 25].

After publishing his work in 1929, Berger wrote 14 original papers about the EEG. Among other findings, these introduced the concept of brain rhythms by describing the existence of the alpha rhythm, also known as Berger wave, as well as the beta wave [26–28]. The application of signal processing techniques for feature extraction and finding of normative values for background rhythms are also among Berger's contributions to the field [29–31].

EEG recording procedures evolved rapidly through the development of innovative electrodes and implementation of amplifiers, filters, active impedance matching and calibration [32–34]. The release of the first commercial electroencephalograph in 1935 allowed the disseminated use of this technique for clinical and research purposes [35]. Less than 30 years after Berger's first publication, EEG recordings were already being used in hospitals [18]. Since then, there have been significant improvements in terms of hardware (amplifiers, digital electrodes, among others), but the same level of development was not seen in EEG analysis, with visual analysis of raw signals remaining the gold standard after a century.

2.2 Signal Acquisition

EEG signals are usually acquired non-invasively, with electrodes placed on the scalp. Abrasion of the scalp and conductive media are used to promote contact between the surface and the electrodes. Invasive EEG recording is also possible, by placing electrodes along the brain's surface [17, 19].

A wide range of electrodes can be used for EEG recording. Needle electrodes are the most common in invasive EEGs, while AgCl electrodes are usually used in the non-invasive variant. Electrode placement on the scalp is usually done according to the 10-20 system (using 21 electrodes) adopted by the International Federation in Electroencephalography and Clinical Neurophysiology in 1958 [36, 37]. For applications that require higher electrode density, different placements are used [38].

The leads from the electrodes are connected to differential amplifiers, which amplify the differences between the inputs, reducing voltage that is common. After amplifying the difference between 1 thousand and 100 thousand times, the signal is filtered. High-pass filtering is used to remove slow artefacts like those related to movement. Aiming to remove higher frequency artefacts such as electromyographic signals, low-pass filtering is employed. Finally, notch filtering can be used to remove the interference caused by the power line (50 or 60 Hz). The EEG signal is then passed through an anti-aliasing filter to prevent information loss and digitized using an analog-to-digital (A/D) converter, usually with a sampling rate between 256 and 1024Hz [39–42].

EEG signals consist of a concatenation of lines corresponding to the plot over time of the differential voltage recorded between a pair of electrodes, which define a channel. The channels in a recording can be set up in different ways, which are referred to as montages. Examples of these are the common average reference montage, where the reference is the average of the outputs of all the differential amplifiers, and the bipolar montage, in which each channel is formed by the difference between two adjacent electrodes [16, 43].

2.3 EEG Signal

2.3.1 Brain Signals

EEGs aim to record the electrical activity of the brain, capturing both physiological rhythmic activity (background) and other transient processes. The resulting signal of a non-invasive recording has an amplitude between 0.5 and 100 μV [16]. An example of a normal EEG can be seen in Figure 2.2.

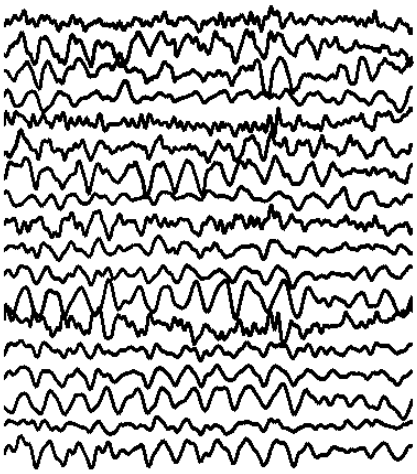


Figure 2.2: Example of a 2 second epoch of a Normal EEG.

Transients are relatively rare events that are not repeated periodically over time. Some of these are physiological, corresponding to normal activity, such as the vertex waves that occur during sleep. Others, like sharp waves and spikes, are associated with pathological events, for instance seizures or interictal activity [44–46].

The rhythmic activity consists of sinusoidal brain waves with a specific frequency. Based on this characteristic, these patterns can be classified as beta waves (frequency between 13Hz and 30Hz), alpha (8-13Hz), theta (4-8Hz) and delta (up to 4Hz). Brain waves from different groups are originated in different regions of the brain and are associated with certain mental or cognitive states. For instance, alpha activity is mostly seen in the posterior part of the skull and it is associated to relaxed states of wakefulness, also being recorded with closed eyes. On the other hand, beta waves are most evident in frontal regions and occur during more active, focused states [47–49].

2.3.2 Artefacts

The presence of corrupting artefacts and interferences in the EEG signal is inevitable. Their origin influences the way they impact the signal, whether it is biological (i.e. related to the patient) or technical (i.e. related to the acquisition) [19].

Muscle activity leads to EMG-related artefacts with relatively high frequencies. Eyeblink artefacts are caused by moving or rotation of the eyeball during blinking. Since there is a potential

difference between the cornea and the fundus of the eye (corneofundal or corneoretinal potential), artefacts occur in the EEG when the eyeball rotates [50, 51]. An example of the effect of muscle activity and blinking on the EEG can be seen in Figure 2.3. Even cardiac activity interferes with the signal, appearing as artefacts that can be mistaken as spikes [52]. The concurrent measurement of biosignals (such as ECG or certain muscle movements) may be useful, since these measurements can be used to visually verify the occurrence of an artefact in the EEG. Technical causes of artefacts include impedance fluctuation, mains interference or issues in the contact between the electrodes and the skin [53].

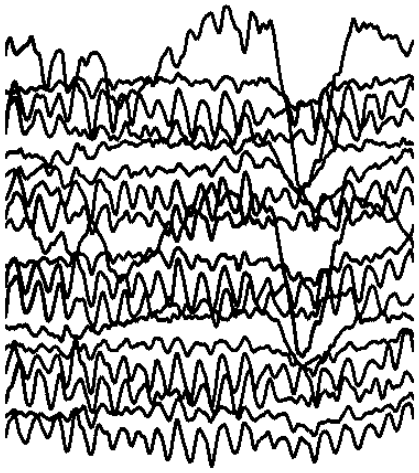


Figure 2.3: Example of a 2 second epoch of a Normal EEG in which it is possible to see an artefact in the last quarter of the signal.

Source separation techniques such as Independent component analysis (ICA) and several variations have been used with the aim of reducing these signal contaminates [54, 55]. More recently, fully automated methods for artefact rejection have been developed [56].

2.4 EEG Analysis

Visual, qualitative analysis of EEG signals by experienced experts remains the gold standard of EEG interpretation [13]. However, this approach is not without drawbacks, ranging from the long training time of the clinicians to the inability of the human eye to fully capture the richness of this signal. Interpreting EEG signals is a time-consuming task, so visual analysis is associated with a high requirement of time and qualified personnel. Furthermore, intra and inter-observer variability reduce the assertiveness of the predictions made based on visual interpretations [14, 15].

To overcome these limitations, quantitative measures derived from EEG signals (quantitative EEG or qEEG) can be used as an alternative way to extract information. This reduces variability, as well as the amount of time and work involved in prognostication [57, 58]. Once the relevant features are defined, methods based on qEEG can be used by non-experts to aid clinicians. Extracted measures may be used alone or in combination, by applying a mathematical model. These

features may include relative delta power asymmetry, wavelet subband entropy, cross-correlation, mutual information, among others [59, 60].

Quantitative EEG shows clear advantages over visual assessment of EEG signals. However, the establishment of relevant features must be done manually and usually involves clinicians. This process is not trivial and it leads to a low efficiency in the development of suitable algorithms [61]. Aiming to solve this, novel approaches such as deep learning are presenting themselves as a possible alternative to further improve EEG interpretation [62].

2.5 Clinical applications

Currently, the EEG plays an important part in research and in the diagnosis of several pathologies, such as depression [63, 64], schizophrenia [65, 66], epilepsy [8, 12] and Alzheimer's disease [67, 68]. Sleep analysis [69, 70] and the monitorization of procedures such as anesthesia [71] are also applications of this technique. It allows continuous monitorization of patients' cerebral activity, characterization of seizures and approximate location of their origin [72, 73].

It is relevant to note that electroencephalography is only one of the techniques used to study and monitor the brain. Others include magnetic resonance imaging (MRI) [74] and its functional variant (fMRI) [75], computed tomography (CT) [76] and positron emission tomography (PET) [77]. MRI and CT are medical imaging techniques that allow us to look at the structure of the brain, while fMRI and PET enable the tracking of cerebral activity and detection of changes [78, 79].

Compared to these techniques, the high temporal resolution of the EEG (in the order of milliseconds) is its main advantage. The direct measurement of brain activity, as opposed to the tracking of indirect markers such as blood flow or metabolic activity used by other techniques, is also a strong point. Furthermore, the equipment used for acquisition is not expensive and it does not require a specific facility. Recordings can be done over a long period of time since the non-invasive variant is painless, and they can even be done in ambulatory. This technique is also safe for the patient since no radiation is used [20, 78].

However, low spatial resolution and the difficulty in reconstructing signal source constitute some of its drawbacks. Since visual analysis is still the most common way of interpreting EEG signals, the time and expertise invested in this task are also limitations. Despite these, electroencephalography is a useful, practical and reliable tool for diagnosis and monitoring that has established itself as standard in hospitals worldwide [18, 78].

Chapter 3

Epilepsy

Epilepsy is a disease of the brain that entails a predisposition to generate epileptic seizures [1, 2]. As defined by the International League Against Epilepsy in 2014, it must satisfy one of the following conditions: at least two unprovoked seizures occurring with more than a day between one another, one unprovoked seizure and a probability of further seizures similar to the recurrence risk after two unprovoked seizures or the diagnosis of an epilepsy syndrome [3]. Due to the variety of epilepsy syndromes and types of seizures, one must consider epilepsy not as a single disorder but as a spectrum of diseases with several causes, symptoms and possible treatments [80, 81].

3.1 Historical perspective

The first description of an epileptic seizure dates from 2000 B.C., in Mesopotamia. Epilepsy was then related to the 'hand of sin', brought about by the God of the Moon. These beliefs continued through Egyptian, Babylonian, Greek and Latin societies [82, 83].

In fact, the word epilepsy comes from the Greek 'to seize, possess or take hold of', since it was believed that epileptics had offended the Goddess of the Moon, and certain positions of the moon melted their brains, leading to madness. Despite this, epilepsy was also considered a 'sacred disease' by the Greeks, synonym of genius, as it affected Hercules and Julius Caesar [84, 85].

Hippocrates, the Greek philosopher, disagreed with both these hypotheses. In his book "On the Sacred Disease", he described epilepsy as a disease of the brain, discrediting divine or wicked origins [86]. He was the first to approach this disorder in a scientific way, proposing possible causes and therapies. Although several Roman physicians shared his belief, the advent of Christianity brought a new era of spiritualism. Epilepsy was connected to witchcraft and led to persecution until the Enlightenment, in the eighteenth century. With the detachment from religion came curiosity and the scientific method, circling back to Hippocrates' hypothesis [85, 87].

Research on the aetiology and therapy of epilepsy continued, accelerating with technological development and availability of techniques such as EEG and MRI [87]. The definition of the disease and of seizures themselves changed throughout the years, according to the available information. In the 1850s, Delasiauve [88] and Reynolds [89] defined epilepsy as a disease without cause

and excluded epileptic seizures of the scope of epilepsy. Gowers, in 1881, once again brought seizures into the definition of the disease [90]. More recently, in 2005, the International League Against Epilepsy proposed an official definition for epilepsy and seizures that was revised in 2014, reducing dubiousness and variability in the diagnosis and communication about the disease [3,91].

Despite these paradigm-changing advances, there is still much to be discovered about the causes, underlying processes and possible treatments of epilepsy [92]. It is also important to note that in countries like Liberia and Swaziland, epilepsy is still linked to witchcraft and possession by spirits. Even in countries where this disease is recognized as a neuronal problem with available therapy, patients still suffer from societal stigma, mostly due to misinformation [82]. This can only be solved by investing in research and improvement of public understanding of the disease, in order to reduce an avoidable 'side effect' of epilepsy for its patients.

3.2 Aetiology

Epilepsy is the fourth most prevalent neurological disorder in the world, affecting population from all age groups and ethnicities [81]. In developed countries, its incidence is approximately 50 per 100 thousand people per year, with more frequent occurrence reported in children and elderly people. This number rises to 100 per 100 thousand people in countries with poor sanitation and inadequate health systems, where the probability of infections is higher [93].

In half of the cases, epilepsy has no discernible cause. In the other 50%, genetic or acquired causes (or a combination of both [94]) can be identified. Epilepsies caused by genetic factors are also referred to as 'idiopathic', while the ones due to acquired factors are called 'symptomatic'. Idiopathic epilepsy is characterized by absence of structural brain lesions and neurological signals, while symptomatic epilepsy is due to some type of identifiable brain lesion [95–97].

Causes for symptomatic epilepsy range from traumatic brain injury to infections in the Central Nervous System, cerebrovascular diseases, brain tumours, degenerative diseases like Alzheimer's disease, developmental disabilities such as cerebral palsy or even febrile seizures [96–98]. In endemic zones where sanitation and health are not ideal, causes like Neurocysticercosis (a parasitic infection of the nervous system) account for 30 to 50% of the cases of epilepsy [99].

3.3 Epileptic Seizures

An epileptic seizure, as defined by the International League Against Epilepsy in 2005, is a transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous activity in the brain [91]. Therefore, these seizures reflect atypical electrical activity characterized by synchronous firing of a large mass of neurons, regardless of the stimuli being excitatory or inhibitory. Epileptic seizures increase the instability of nerve elements, facilitating further occurrences [100, 101].

A seizure type entails a unique pathophysiological mechanism and it may be associated with a specific cause, having its own prognosis and adequate therapy. This is more specific than an

epilepsy syndrome, which is a group of signs and symptoms that define a unique epilepsy condition, imperatively involving more than one seizure type [95, 102].

The classification of epileptic seizures and syndromes is dynamic, reflecting the growing knowledge of the underlying physiology of the disease. Currently, the proposed classification takes into account the type of seizure, whether they are focal (i.e. affecting only part of the brain) or generalized (i.e. affecting both hemispheres of the brain), the syndrome, cause and associated deficits [93, 95].

Seizures, also known as ictus or the ictal state, are often preceded by the aura [103]. In this earlier phase, the emotional state of the patient and their senses such as smell and taste may be altered [104–106]. The ictus itself may result in loss of consciousness (common in generalized seizures), convulsions, spasms, as well as unfamiliar behaviours and sensations [107–109]. After a seizure, in what is known as the post-ictal state, the patient may experience confusion, dizziness, drowsiness, blurred vision or ataxia, among other symptoms. The post-ictal state usually lasts between 5 and 30 minutes, but for some patients it may be longer, further hindering their recovery [110, 111]. The seizures may also have longer-lasting consequences such as trauma, burns and bleeding [109].

A particularly dangerous type of seizure is status epilepticus (SE). It is defined as an epileptic seizure lasting for more than 5 minutes or several seizures within 5 minutes without return to the pre-convulsive neurological baseline [112]. Prolonged and repetitive seizures are less likely to end spontaneously (i.e. without therapy administration) and they have been linked to irreversible brain damage and pharmacoresistence. Therefore, the clinical prognosis in cases of SE is worse than for other types of epileptic seizures and the mortality is higher [113–115].

In general, the mortality of epileptic patients is increased, with sudden unexplained death in epilepsy (SUDEP), suicide, status epilepticus and the effects of the seizures as some of the causes [107, 116]. Aside from mortality, epilepsy has other consequences of neurobiological, cognitive, physical, psychological and social nature that impair the quality of life of the patients. Thus, reducing the frequency of the seizures is of utmost importance to reduce the impact of the disease [92, 117]. This can be done through early, assertive diagnosis and adequate treatment, which will be addressed in the following sections.

3.4 Diagnosis

3.4.1 Misdiagnosis

Epilepsy encompasses a plethora of syndromes and types of seizures, some of which are similar to abnormalities associated to other diseases. Thus, distinguishing a non-epileptic paroxysmal attack from an epileptic one is not without doubt and the rate of misdiagnosis for epilepsy is high. It is estimated that about 30% of patients diagnosed with epilepsy actually suffer from another condition [4, 5].

First seizures, which are not synonym of epilepsy according to its definition, sometimes lead to misdiagnosis. Other conditions such as diabetic seizures, nonepileptic seizure disorders, such as Tourette Syndrome and narcolepsy, meningitis, some cardiac diseases or even eclampsia during pregnancy are often confused with epilepsy [4, 118].

This entails increased risk, as the patients are not being treated for the disease they have and are sometimes given medication with potentially harmful side effects that may worsen their condition [6, 7]. Thus, proper diagnosis is of extreme importance both to reduce the frequency of seizures and to prevent dangerous consequences of misdiagnosis.

3.4.2 EEG as a diagnostic tool in epilepsy

EEG is currently the most useful technique for this task [1, 8]. Ictal EEG, i.e. EEG measured during a seizure, is the only method that nearly always unequivocally distinguishes an epileptic seizure from a non-epileptic one, allowing certain diagnosis of the disease. It also aids in the identification of the source and type of seizure, facilitating therapy administration. However, the likelihood of acquisition of an ictal EEG is low due to the unpredictability of its occurrence [9, 10, 119].

Recording of interictal EEGs, i.e. brain signals from prospective patients in periods where no seizure is happening, is always possible and therefore widely used to aid in diagnosis [11, 12]. Interictal Epileptiform Discharges (IEDs) are transient patterns that help to differentiate epilepsy from other conditions. Epileptiform patterns include mainly spikes and sharp waves, shown in Figure 3.1, which are characterized by high amplitude and short duration (20 to 70 ms for spikes and 70 to 200 ms for sharp waves). These patterns correspond to paroxysmal depolarization shifts and are usually followed by a slow wave lasting 200 to 500 ms, linked to hyperpolarization. IEDs are present in about half of the recorded EEGs from epileptic patients and this number rises to 80% with repeated recordings. EEGs recorded during sleep are more likely to include IEDs since their incidence is higher in this state [1, 87, 120].

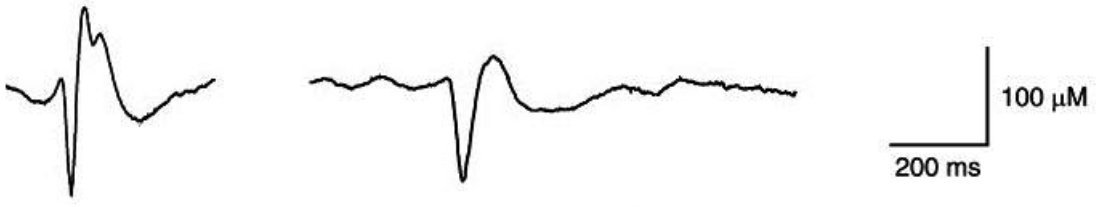


Figure 3.1: Interictal patterns: on the left, an interictal spike; on the right, a sharp wave [120]. It is possible to see their high amplitude (compared to the normal amplitude of the EEG signal), as well as short duration.

While these patterns show evidence of abnormal cortical hyperexcitability and hypersynchrony during a seemingly asymptomatic state, they are not enough to diagnose epilepsy, since normal subjects or patients suffering from other diseases can have EEGs where IEDs are found [12, 61].

Despite this, the presence of IEDs can aid in the diagnosis of epilepsy and information derived from the EEG such as the frequency of the spikes and the location of their origin in the brain can be insightful in what concerns the determination of the epileptic syndrome [8].

Other uses of EEG in regards to epilepsy include monitoring of SE, assessing efficacy of the prescribed therapy and choosing eligible patients for epilepsy surgery [12, 121]. Aside from EEG, imaging techniques like MRI or CT (when access to MRI is restricted) may help detect changes that could underlie refractory focal epilepsies and assist in electing patients for surgery. While visual analysis remains the gold standard for these techniques, automated analysis has become an important aid and, with increasing research on this field, it may make diagnosis more efficient by reducing time and increasing assertiveness [8, 93].

3.5 Treatment

Adequate therapy to reduce seizures is paramount for a better quality of life of epileptic patients. Anti-epileptic drugs (AEDs) are currently the first line of therapy for epilepsy. When one or a combination of these drugs is not effective, the epilepsy is said to be refractory. In these cases, non-pharmacological therapy is available [83, 87, 122].

3.5.1 Anti-Epileptic Drugs

AEDs can be defined as preventive chemicals that reduce neuronal synchronicity to avoid seizures. They are effective in up to 70% of epileptic patients, although some patients have to try different AEDs before finding one of more drugs that reduce seizure frequency [117, 122].

Historically, the first therapy for epilepsy, bromides, was discovered in the mid-nineteenth century by Charles Locock. It was widely used until phenobarbital and phenytoin were discovered in the beginning of the twentieth century and became the standard treatment for epilepsy. Until the 1990s, sodium valproate, carbamazepine, primidone and ethosuximide joined the range of available AEDs. This group of compounds is generally referred to as ”old drugs”, as opposed to the ”new drugs”, discovered after the 1990s. New drugs include tiagabine, pregabalin, gabapentin, topiramate, clobazam, oxcarbazepine, vigabatrin, lamotrigine and levetiracetam [87, 123].

Currently, there are over 20 different drugs licensed for treatment of epilepsy, with different mechanisms of action and aiming to treat different types of seizures or syndromes. Although not all mechanisms of action are well understood, some have been widely studied. For instance, carbamazepine and phenytoin block sodium channels while tiagabine and vigabatrin, among others, work by enhancing the inhibitory GABAergic system [93].

Both new and old drugs continue being used, with no significant difference in effectiveness being recorded. Older AEDs entail lower costs and are more widely available, but newer drugs usually show lower levels of toxicity. Choosing the appropriate therapy for the patient’s syndrome is more important than prescribing a ’new’ or ’old’ drug since, if this choice is not correct, the condition may be aggravated by AEDs. For instance, if a patient has seizures due to inhibitory synchronous activity and is prescribed an anti-epileptic drug that increases inhibition or decreases

excitation, it is likely that the frequency of the seizures will rise [93, 122]. It is also important to take into account patient-specific characteristics such as age, sex and medication prescribed for other conditions. Additional care needs to be taken in cases of female patients taking oral contraceptives, since AEDs may reduce its effectiveness and the oestrogen in the contraceptives could lead to recurrence or exacerbation of seizures [124, 125].

Aside from this, guidelines state that monotherapy (i.e. treatment with only one AED) should be administered at the lowest effective dosage to make the patient seizure-free. If monotherapy is not effective after trying several drugs, polytherapy (i.e. treatment with a combination of AEDs) may be considered. This is not ideal since it increases the probability of poor compliance, drug interactions, teratogenicity and toxic effects. If polytherapy is employed, the chosen drugs and the dosages should be such that minimize interactions and side effects, maximising synergy [122, 126, 127].

Despite all these indications and guidelines, up to 50% of epileptic patients on monotherapy with AEDs experience side effects. These include fatigue, drowsiness, dizziness, blurred or double vision, headaches, impaired motor skills, memory or concentration. Treatment with AEDs may also lead to rashes, hematologic dyscrasias, hepatotoxicity, bone density loss, gingival hyperplasia and neuropathy [109, 123, 126]. If these side effects cannot be eliminated by using a different AED or polytherapy, or if a significant (or total) reduction in seizure frequency is not possible, one must resort to non-pharmacological approaches [87, 122].

3.5.2 Non-pharmacological therapy

Non-pharmacological therapy, under the form of epilepsy surgery or vagus nerve stimulation (VNS), can be used in cases of refractory epilepsy. Approaches such as deep brain stimulation, other types of neurostimulation, cooling, optogenetics and dietary treatments such as the ketogenic diet are being studied as possible therapeutics for this disease [87, 93, 128].

Epilepsy surgery may be performed in cases where the area responsible for the seizures can be determined and is limited to a particular non-eloquent region. If the whole brain is identified as responsible for the seizures, epilepsy surgery is not an option. However, in some patients, it is possible to identify a trigger area, as is the case in some generalized 'thalamo-cortical' epilepsies. The identification of the origin of the seizures is usually done using EEG recordings. When non-invasive EEG is not enough to identify the area, invasive EEG techniques may be employed. Depending on the case, either resective or non-resective surgery can be performed. In resective surgery, the origin of the seizures is removed, while in non-resective surgery there is a physical separation between that area and the rest of the brain, without removal [129, 130]. In general, resective surgery leads to a higher probability of the patient becoming seizure free. The procedure with the highest success rate is the temporal lobe resection, with 70% of the patients reporting a seizure-free life [93].

In cases of generalized refractory epilepsy or when patients fail to qualify for surgery, VNS can be used to potentially reduce seizure frequency. A pulse generator is implanted on the patient and connected to the vagus nerve, in the neck. By mildly stimulating this nerve regularly, it is possible

to reduce the irregular synchronicity and thus reduce seizures. The success of this method is largely dependent on the patient, but it has shown to reduce seizure frequency in up to 50% in 30 to 40% of the patients [131, 132]. Deep brain stimulation has recently received more attention as a possibility of treatment for refractory epilepsy when surgery is not an option.

Despite the effectiveness of both pharmacological and non-pharmacological approaches, many patients still have recurrent seizures or suffer from side effects of the prescribed therapies [124]. Thus, incessant research in this field is needed to reach more patients and improve their quality of life.

Chapter 4

Machine Learning

Machine learning can be defined as the set of computational methods that use data or experience to improve performance on a certain task, generalizing from examples [133–135]. Deep learning is an innovative subfield of machine learning that encompasses a set of techniques and methods inspired by the human brain and its learning processes [136].

Using machine learning methods, it is possible to create a useful approximation of reality by taking data regarding a problem and creating an algorithm. [134, 137]. Deep learning does this with artificial neural networks that learn from data using several layers with increasing levels of abstraction. Since the network itself is responsible for the feature extraction process, it becomes almost independent of human knowledge, which reduces the time needed to develop an algorithm as well as the field expertise needed to do so [136, 138, 139].

4.1 Historical Perspective

The history of deep learning and machine learning is closely related to that of artificial intelligence and pattern recognition. It is also inevitably intertwined with several other areas of knowledge such as computer science, physics, mathematics, statistics, logic, philosophy and cognitive neuroscience [133, 140].

The will to 'create intelligence' can be traced back to antiquity, where men wanted to 'forge the gods' [141]. However, it took years of scientific progress for the modern concept of Artificial Intelligence to develop [142]. Developments such as the mechanical adder, the binary system or Boolean logic, culminating in the invention of the computer in the 1940s, allowed substantial advances in this area of research. In the 1950s, there was a growing interest in computational approaches to learning, as learning was identified as a central part of intelligent systems and it became possible to join that with computational power [143–145].

4.1.1 The Perceptron

The first general purpose algorithms were in the scope of neural modeling and decision theory. The groundwork for this paradigm derived from mathematical biophysics, with Rashevsky [146]

and McCulloch and Pitts [147] translating neural activity into propositional logic, allowing its computational modeling. In 1962, Rosenblatt presented a model of an artificial neuron called the perceptron. Its inputs (x_0 to x_n, with x_0 being the bias, equal to 1) were combined with varying weights (w_0 to w_n) as can be seen in Figure 4.1. This resulted in a weighed sum given by $\sum_{i=0}^{n} x_i * w_i$ that was then passed through a step function, predicting 1 if the result was above a certain threshold (influenced by the bias) and 0 otherwise [148, 149].

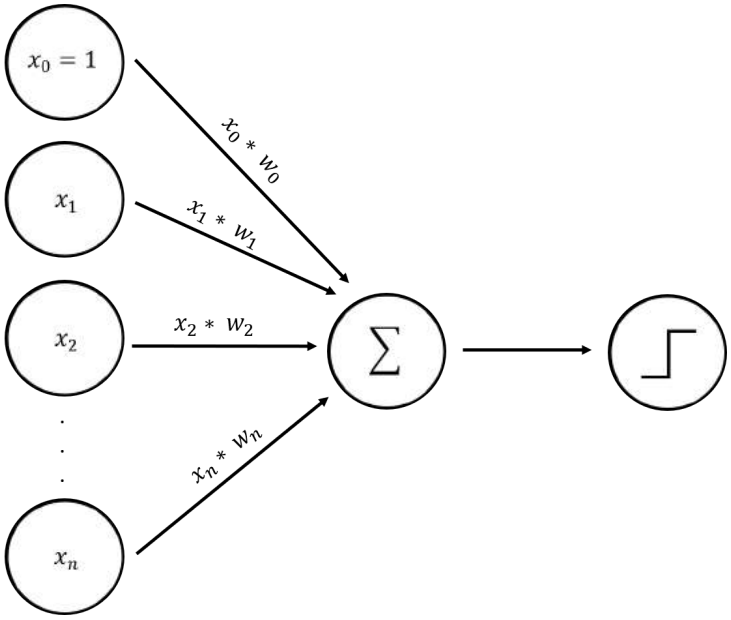


Figure 4.1: Basic structure of a perceptron, as described by Rosenblatt.

This was a moment of glory for connectionists, the researchers that believed that a universal learner could be achieved by modeling neural phenomena with neural networks, as the perceptron was able to mimic human neurons in a concise way and solve linear classification problems with a simple algorithm. Concurrently, other types of algorithms, like those based on the simulation of evolutionary processes, started gaining traction among the machine learning community. Alternative but powerful approaches included the use of statistical decision theory [150–152] and the development of discriminant functions based on a group of examples, of which the most popular example is Samuel's checkers program [153]. Other methods were based on logic and graph structures instead of statistics and mathematic, using inverse deduction and manipulating symbols to acquire knowledge [154–156].

As can be concluded from the diverse approaches to machine learning, the 1960s were prolific times for this field. However, this exponential growth was halted in the mid-seventies, following the publication of 'Perceptrons' by Minsky and Papert in 1969 [157]. In their work, Minsky and Papert showed that perceptrons were not able to solve problems involving non-linear spaces and thus could not be used to model problems as simple as the XOR function. As this proved that the perceptron was not the universal learner that it initially aimed to be, connectionism was almost completely abandoned. Further delays in research were caused by the disillusion in artificial

intelligence and subsequent cuts in funding by the British and American governments [158, 159].

4.1.2 The Multilayer Perceptron and Backpropagation

Some work on linear models kept being developed, but no ground-breaking discoveries were made until 1986, when the backpropagation algorithm was rediscovered by Rumelhart [160, 161], after it was first published by Werbos in 1974 [162, 163]. Backpropagation substituted the McCulloch and Pitts model and allowed the organization of networks of interconnected neurons.

Multilayer Perceptrons (MLPs), feedforward neural networks comprised of interconnected neurons grouped into layers, became feasible with backpropagation. Before, this was not possible because there was no way to derivate the error with more than one layer. MLPs included an input and output layer and at least one layer in between (i.e. at least one hidden layer). The neurons of these layers could have any activation function, but non-linear functions were usually used for this purpose, as they prevented the system from collapsing to a linear modeling and allowed it to learn more complex decision boundaries [164].

Backpropagation calculated the partial derivative of the cost function with respect to each weight (i.e. the gradient), repeating this process backwards in the network. After calculating the gradient for all layers, ending in the first layer, the weights were updated according to the value of the gradient and to the defined learning rate, a small constant used to avoid large steps in the update. The equation for weight update can be simplified as new weight = old weight - gradient * learning rate, which indicates that positive gradients lead to a reduction in weight and vice versa, making weights converge to a value that minimizes error [160]. This gradient descent algorithm was quite efficient since it used this backward flow to calculate the value for the previous layers instead of computing it from scratch.

This widened the problems that could be solved using connectionist algorithms, relaunching research in the area. Aside from MLPs, non-linear extensions to generative linear models were also developed, along with other algorithms like regression trees [164, 165].

The Support Vector Machine [166] was one of the most important developments in machine learning after backpropagation. The algorithm generalized from similarities in the training data to make predictions, knowing that non-linear feature spaces could be mapped to higher dimensions, where the boundary between them was linear and learnable by this vector machine.

4.1.3 Long Short-Term Memory networks

A crucial development in connectionism was the Long Short-Term Memory (LSTM) network. LSTMs are a type of recurrent neural networks (RNNs), which are cyclic graphs, unlike feedforward networks. Also, while feedforward networks map one input to one output, RNNs can have more than one input or output (or both). RNNs possess 'memory', which is able to store previous information in states and use it to aid predictions, making them useful in handwriting or speech recognition [167–169]. For instance, describing an image through a string of words is mapping one input to many outputs, while translation is an example of multiple inputs and multiple outputs.

Traditional RNNs have some issues dealing with long-term dependencies, along with vanishing and exploding gradients. LSTMs were developed by Hochreiter and Schmidhuber to deal with these issues [170]. They are made up of a chain of layers, similarly to traditional RNNs, but instead of repeating a single network layer, the cells are composed of four layers that work together to decide what information to keep, how to update the state of the cell and to produce an output for the following cell (see Section 4.3.2.1 for more details).

4.1.4 ImageNet

In the following years, interest in machine learning continued to rise after IBM's Deep Blue defeated chess champion Garry Kasparov in 1997 [171]. In 1998, LeCun released the MNIST database of handwritten digits, allowing researchers to use the same data and thus directly compare results of different methods. In the same year, LeCun proposed LeNet-5, a convolutional neural network (CNN) to automatically classify the MNIST digits [172]. CNNs are feed-forward neural networks that use convolution operations to extract features, and they will be further discussed in Section 4.3.1.

ImageNet, a large database that currently includes over 14 million labeled images in more than 20 thousand categories, was created in 2009 [173]. To boost the use of this database, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was created in 2010. It consisted in using a subset of the ImageNet database to train a machine learning algorithm, aiming to surpass human accuracy in image classification.

In 2012, the winner of the ImageNet Competition was AlexNet, developed by Alex Krizhevsky [174]. AlexNet was a CNN that included 5 convolutional layers with ReLu activation, pooling and dropout layers and a Softmax with 1000 units, optimized by a batch stochastic gradient descent optimizer. It took five to six days to train on two Graphical Processing Units (GPUs) and it achieved 16.4% top 5 error, against the 26% yielded by the winner of the previous year. This innovative model is said to have been the beginning of the AI boom of the 2000s [175]. Interest in machine learning, and in neural networks in particular, peaked, as did investment in the field. The wider availability of GPUs, circuits that sped matrix multiplication, leading to a faster training process, also allowed heavier architectures and more innovation in the neural networks used.

The VGG Network, developed in Oxford by Karen Simonyan and Andrew Zisserman in 2014, was the runner-up in that year's ILSVRC [176]. It used smaller filters than the AlexNet and its architecture was deeper, taking two to three weeks to train on 4 GPUs. Although it did not win the competition, its flexible architecture led to vast use in the field.

The winner of 2014's ILSVRC was GoogLeNet, proposed by Szegedy and his team at Google [177]. It introduced the Inception module, which consisted in using parallel filters of different sizes to capture different patterns that were stacked in a feature map. Convolution with 1x1 filters was used to avoid dimensionality increase within the modules. Using several of these modules to create a wider network, GoogLeNet managed to decrease top 5 error to 6.7% and increase computational efficiency. The team continued to improve this model over the years, leading to several versions of the now named Inception network.

In the following year, Microsoft's ResNet (Residual Network) won the ImageNet competition with a top 5 error of 3.6%, under the 5% achieved by humans [178]. The two main innovations of the ResNet were its depth (152 layers) and the residual module. In fact, it is the use of the residual that allows networks to have such depth without degradation and vanishing gradients. These problems arise because, during backpropagation, repeated multiplication makes the gradient increasingly small, leading to higher training errors when depth is continuously increased. Assuming a set of connected layers have as input x and yield a function $H(x)$, using a residual function defined as $F(x) = H(x) - x$, it is possible to optimize the residual instead of the unreferenced mapping without adding parameters or increasing complexity. Both functions approximate the same target, but the residual does it more effectively due to its formulation, solving degradation. Shortcut connections were used to perform the identity mapping, carrying information from previous layers forward in the network.

Since the goal of surpassing humans had been reached, the ImageNet Competition stopped after 2017. Other relevant achievements in the field include Facebook's DeepFace project [179], which was able to identify human faces with over 97% accuracy and Google's AlphaGo project, which was able to defeat the Go champion in 2016. This algorithm was improved in 2017 into AlphaZero, which was additionally specialized in other two-player games, including chess [180, 181].

4.2 Types of Learning

Although it is said that machine learning, and, consequently, deep learning algorithms 'learn from experience', this is not enough to explain the learning paradigm involved. The data used for training, the type of learning and the learning task at hand are crucial when choosing an algorithm and largely influence learner performance [182].

Data is one of the most important factors since it is used to train the learners and, as such, it has a great impact on their performance. The data used to train algorithms is just a sample of the real-world data, so volume (i.e. how much data is available), representativeness (i.e. how diverse it is in relation to the real-world data) and quality (related to how noisy or omissive it is) are some of the crucial characteristics that must be taken into account when choosing and training learners [133, 183]. For deep neural networks in particular, the volume of data is crucial because the layers rely solely on raw data to learn. For lower volumes, other machine learning algorithms such as SVMs may achieve better performance.

The amount of information concerning the true class or value of each data sample is also of paramount importance, since it influences the learning paradigm. Labeled data is data for which the true class is known, while, for unlabeled data, the true value is not available [183].

Over time, machine learning has branched into different ways of dealing with learning, depending on the task at hand and on the available data. The taxonomy of learning paradigms is not absolute, since there are several distinguishing characteristics that can be used for this classification. For instance, learners can be classified as active or passive, according to their interaction

with the environment. A passive learner can only observe the provided information while an active learner can pose queries or perform experiments during training. Another possible distinction is between batch and online learning. In batch learning, a model is built based on the available data and it is used to make predictions. On the other hand, in online learning, the model is updated upon the success of each interaction with the environment [182].

The degree of supervision during learning is one of the most commonly used ways to classify the type of learning [184].

4.2.1 Supervised learning

Supervised learning is used when there is a dataset that includes the information needed to create a model of the problem (labeled data). The algorithm looks at this information, builds the model and, when presented with new data, it should be able to generalize and respond correctly [183, 185].

The main tasks that can be solved with supervised learning are regression and classification [184]. Regression predicts a numerical value for each data point, while classification aims to predict a discrete class label for each new instance. In some cases, classification works with continuous values, similarly to regression, but then discretizes them into classes.

4.2.2 Unsupervised learning

Unsupervised learning means finding similarities within the provided data to try to model its structure, given that there is not enough information regarding the data to directly build a model (unlabeled data) [184].

Association, clustering and dimensionality reduction are the types of tasks usually tackled by unsupervised learning [182, 184]. Association aims to determine the co-occurrence of events, while clustering groups instances through a measure of similarity. When a new instance is presented, it is assigned the class of the most similar cluster. Finally, dimensionality reduction consists in reducing the number of variables while keeping its discriminant characteristics. This can either be done through feature selection, which chooses the most distinguishing subset of variables or through feature extraction, which consists of transforming the variable space into a one with lower dimensionality.

4.2.3 Semi-supervised learning

Semi-supervised learning is used when there is a large amount of unlabeled data and a smaller amount of labeled data. This paradigm aims to use both types of data, surpassing the performance that could be obtained with either supervised or unsupervised learning. To do that, it is necessary to make assumptions about the data distribution, whether it is regarding its continuity, clustering or others [182].

4.3 Deep Learning Models

Deep learning uses artificial neural networks (ANNs) that derive from Rosenblatt's perceptron, and thus are inspired in the brain and its processes [136, 186]. As discussed in Section 4.1, there are several types of ANNs, from which CNNs and LSTMs, a type of RNNs, can be highlighted.

4.3.1 Convolutional Neural Networks

CNNs are feedforward networks with layers entailing increasing levels of abstraction. These layers are made up of interconnected neurons with learnable weights and biases. The sequence and parameters of the layers constitute the architecture of the model, which determines what the network learns from the data. The optimizer and the loss function are also key determinants in how the network learns, affecting performance [136, 186].

4.3.1.1 Layers

Each network contains visible layers (i.e. the input layer and the output layer) and hidden layers. The input layer nodes receive a single value and pass them to the first hidden layer. Hidden layers transform data so that it can be used as an input for the next layer. The set of hidden layers in a CNN can be divided into a convolutional and a classification subset. In traditional CNNs, hidden layers are stacked linearly, while architectures such as Google's Inception showcase non-linearly stacked layers (inception module) [177]. Finally, output layers transform the output of the last hidden layer to a range or shape that is meaningful for the problem.

Convolutional Layers

Convolutional layers are a set of filters (kernels) that slide across the input to detect if a particular pattern is present. Convolution, an element-wise product and sum between the filter and the input matrices, is used to perform this operation. While the kernel size of each filter (i.e. its receptive field) is a hyperparameter, they extend through the full depth of the input, generating a 2D activation map at the end of a forward pass. Since there is a set of filters being used, the stacking of the activation maps yields a 3D matrix [186].

Stride and padding are two other important hyperparameters of convolutional layers, responsible for the control of the spatial size of the output of the layer. Stride represents the shift made by the filter during the forward pass. This means that higher strides result in smaller outputs. If the stride is bigger than one, it may be necessary to perform padding, usually adding zeros around the border of the input (zero-padding), to ensure that the filter 'falls' within the input.

As mentioned in Section 4.1.2, the use of non-linear activation functions is paramount for the stacking of layers. These allow networks to model virtually any function, approximating the data more accurately. Historically, the Sigmoid function given by $1/(1+e^{-x})$ was widely used for this purpose, mapping the output to an interval between 0 and 1. However, since there is saturation for large values of x, the gradient becomes very low and the problem of 'vanishing gradient'

arises. To solve this issue, Hinton [187] presented the Rectified Linear Unit (ReLu) function, given by $max(0,x)$ [187]. Since there is no saturation of the gradient, vanishing gradients do not occur. Also, this function is less computationally expensive and it promotes sparse activation of the neurons, since it yields 0 for all neurons where $x < 0$.

Pooling Layers

Pooling layers are usually used in between convolutional layers, aiming to reduce the size of the representation, and thus the number of parameters (weights) to be learned, while also reducing overfitting. Pooling is usually done by taking the maximum value in a certain window (max pooling), but it may be done employing an averaging function or others. Since it makes the network focus on a smaller number of neurons, it has a regularizing effect, improving the generalization power [188].

Dropout Layers

Another way of reducing overfitting is the use of dropout layers. These can be used after pooling or between fully connected layers. Dropout consists on 'killing' (i.e. not activating) a user-defined percentage of the neurons, chosen at random, on each presentation of a training case. By averaging the weights in the end of this process, one obtains a more general model [188].

Fully Connected Layers

Fully connected layers are responsible for producing the output. These can only deal with one-dimensional data, so a flattening layer is needed between the previously described layers and the fully connected ones. Flattening transforms the 3D stack of activation maps into a 1D vector. Neurons in fully connected layers have connections to all the neurons from the previous layer, behaving like traditional multi-layer perceptrons [186]. The final layer must have the same number of units as the classes in the output, so a CNN for binary classification would have 2 units in its last layer. In a binary case, the Logistic function can be used to yield the probability of each class. The Softmax, a generalization of the Logistic function, is usually used in the last layer to perform multi-class classification.

4.3.1.2 Optimization

Training CNNs entails high computational costs. Optimization techniques have been developed, aiming to lower these costs and make training more efficient. In the convolutional layer subset, the large amount of operations slows down training. The high number of parameters in the fully connected layers further contributes to a longer training time. Optimization aims to increase the effectiveness of the maximization of an objective function (or, inversely, the minimization of a cost or loss function), which is a function of the network's parameters [189, 190].

Gradient descent, described in Section 4.1.2, calculates the gradient for the entire batch of data and performs an update [160]. For large datasets, this can result in a very slow optimization and, for non-convex surfaces, it may also lead to convergence in a local minimum instead of a global one. To solve these problems, some variations such as the stochastic gradient descent (SGD) and the mini-batch gradient descent, can be used. While these reduce the time needed to converge in

a minimum, the choice of a learning rate and the occurrence of local minima remain issues of this class of methods [189].

There have been several methods that build on gradient descent, the most successful one being Adam (Adaptive Moment Estimation) [191]: $\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\bar{v}_t}+\varepsilon}\bar{m}_t$, where $\bar{m}_t = \frac{m_t}{1-\beta_1^t}$ and $\bar{v}_t = \frac{v_t}{1-\beta_2^t}$.

As can be seen in the formulas above, it uses the first and second moment of the gradient ($\bar{m}_t$ and $\bar{v}_t$) to adapt the learning rate, η, for each parameter. It calculates the moving average of $\bar{m}_t$ and $\bar{v}_t$ and uses β_1 and β_2 to control the decay rates. ε is used to prevent divisions by zero. Adam makes convergence faster, reduces fluctuations in parameter values and avoids the vanishing learning rate problem that arises from using only the first moment to update it.

4.3.2 Recurrent Neural Networks

As described in section 4.1.3, RNNs are cyclic directed graphs like that of Figure 4.2, which take into account the present input but also past inputs to make decisions.

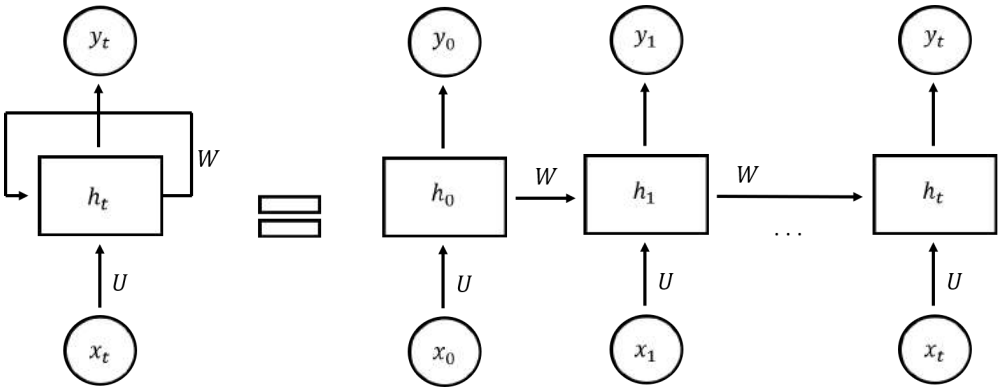


Figure 4.2: Possible architecture of a RNN. It is possible to see that a module is repeated, being applied to the past and present inputs.

Past information is kept on the network's hidden state (h_t), given by $h_t = \phi(Wx_t + Uh_{t-1})$. Thus, h_t is a function of the current input, x_t, and of the previous hidden state, h_{t-1}, multiplied by weight matrices (W and U) [192]. These matrices are the weights that determine how important the present and past states are, and they are used to minimize the error using an algorithm called backpropagation through time (BPTT) [162]. The mapping function is ϕ, which can be a logistic function, making gradients manageable by BPTT.

This algorithm has a similar principle to backpropagation, but it calculates the errors for each time step, accumulating them. The update of the weights is done in the end, given that W and U are the same throughout the network. This process is repeated until the error is minimized. BPTT becomes slow when there are many time steps, since there is a hidden unit per time step. Sometimes, a high number of time steps is necessary for longer persistence in memory, but it may make the network very computationally expensive. Newer algorithms such as the Truncated BPTT allow processing of a pre-determined amount of time steps, reducing training time [193].

4.3.2.1 Long Short-term Memory

LSTM units can be used as building blocks of an RNN, constituting LSTM networks, mentioned in Section 4.1.3. They have a more constant error (i.e. the error has a smaller variation between training epochs) than traditional RNNs (such as the previously described ones that entail the repetition of a single module with a simple structure, similar to a layer of a feedforward network), allowing the use of more time steps [194].

Each LSTM unit has four layers: a memory cell, an input gate, an output gate and a forget gate [195, 196], as can be seen in Figure 4.3.

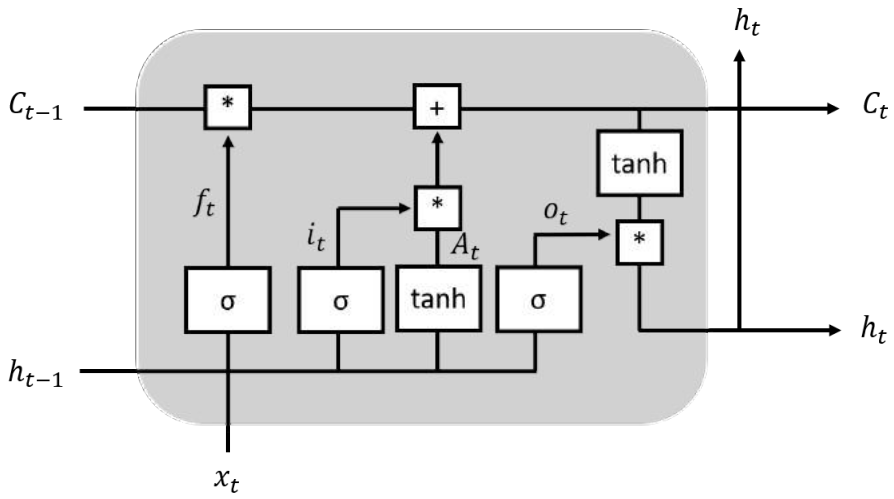


Figure 4.3: Representation of a LSTM unit. The forget gate (first vertical line on the left) decides how much of the previous hidden state, h_{t-1}, should be kept according to f_t. The input gate (second vertical line) performs an analogous decision, given by i_t. The cell state is updated from C_{t-1} to C_t after the operations represented by the third vertical line. The output (last vertical line) is given by the multiplication of the value at the output gate, o_t, and the result of the application of the hyperbolic tangent to the updated cell state C_t.

The cell state carries the flow of information through the network, with the cell state being changed in each unit by its layers. The forget gate decides how much of the previous hidden state, h_{t-1} should be kept according to $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$, which is also a function of the current input x_t. The sigmoid maps the result between 0 and 1, with 1 meaning that all the information from the previous hidden state is kept. The input gate performs an analogous decision, given by $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$. Concurrently, a vector of candidate values to add to the cell state is created according to $A_t = tanh(W_c[h_{t-1}, x_t] + b_c)$. A hyperbolic tangent function is used in this step, with the same purpose as the sigmoid in the previous ones. The cell state is then updated using $C_t = f_t * C_{t-1} + i_t * A_t$. The output of each unit is given by the multiplication of the value at the output gate, calculated through $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$ and the result of the application of the hyperbolic tangent to the previously calculated cell state. The resulting formula is $h_t = o_t * tanh(C_t)$. This output may exist for every unit if the mapping done by the network is

one to many or many to many. However, it is possible to have outputs for some units or just for one of them, if the mapping is one to one or many to one.

It is important to note that there are variants of the traditional LSTM, such as the peephole LSTMs or LSTMs with coupled forget and input gates [195].

4.4 Performance Estimation

4.4.1 Metrics

To assess the performance of an artificial neural network on a given dataset, a range of metrics can be used, depending on the problem. Accuracy is a common metric for this purpose, as it calculates the rate of true classifications: (TN+TP)/(TN+TP+FP+FN), in which TN is true positive, TN is true negative, FP is false positive and FN is false negative. However, the choice of this metric is not trivial since some problems may not have the same cost value associated with erring a positive or negative observation [197]. In these situations, metrics such as sensitivity (TP/(TP+FN)), specificity (TN/(TN+FP)) or precision (TP/(TP+FP)), among many others, may be better suited to assess performance.

Performance estimation may also be done using graphical methods such as the Receiver Operating Curve (ROC curve), created by plotting the sensitivity against the false positive rate, given by 1-specificity for several thresholds. The area under this curve (AUC) can also be used as a performance measure.

4.4.2 Overfitting

The values of the metrics described above are not enough to assess the performance of an algorithm. The data that is used in their calculation is of the utmost importance, since the learner will usually demonstrate better performance on the data it has already seen (i.e. training data) than on unseen data. If this difference in performance is very large, the model is said to have high variance, which means it is learning spurious patterns from the training data instead of learning relevant features that it can generalize when presented with new data. This phenomenon is also known as overfitting [198].

Overfitting may be reduced using several techniques, including an appropriate choice of architecture and optimizer, which have been previously discussed. The use of regularization (complexity reduction) under the form of dropout or pooling layers further contribute to lower model variance. Adding a regularization term on the norm of the weights is also an option to reduce overfitting since it penalizes complexity in the model [199]. Batch normalization, which is done by normalizing the inputs and scaling the activation, also contributes to this end goal [200].

If the volume of training data is not large enough, data augmentation can be used to increase it and reduce overfitting. Data augmentation consists in creating new data based on the available data, allowing the network to look at more examples and becoming more robust. It is worth pointing out that data augmentation is only applied to the training dataset. Simple techniques

for this include flipping the inputs vertically or horizontally, translating objects, rotating, scaling, cropping or adding gaussian noise do the input [199].

More advanced techniques for data augmentation involve the use of Generative Adversarial Networks (GANs) or neural style transfer. GANs are a set of two neural networks, in which one produces an artificial input and the other evaluates it to see if it is a good enough forgery of the available training data [201]. Neural style transfer consists in updating the input instead of the weights, to match a style (captured by a set of parameters) from another input [202].

4.4.3 Cross-validation

Cross-validation includes a set of methods that allow performance estimation in unseen data when a separate set of data is not available for testing. The simplest form of cross-validation is the holdout method, which is the splitting of the available data into two sets, one used for training and another one for testing [203]. This way, by calculating metrics for both sets, it is possible to assess the occurrence of overfitting.

K-fold cross-validation is another way of performing cross-validation, and it is particularly useful when the available data is not enough to split into training and testing. Unlike the holdout method, all the available data is used to train the algorithm. This is done through an iterative process where the data is randomly shuffled and partitioned into k folds, with one of them being used for testing and the rest (k-1) for training, changing the testing fold in each iteration. The algorithm's performance can be averaged over these folds, leading to a more accurate estimate of the training and test error, as well as overfitting [203]. This may be computationally expensive since the algorithm is trained from scratch k times. Leave-one-out cross-validation is a particular case of k-fold cross-validation, where k is equal to the number of points in the dataset.

4.5 Visualization

Although deep neural networks have shown great potential across areas, these algorithms are not without shortcomings. Aside from the high computational cost and training time, alleviated by optimization, and the possibility of overfitting, reduced with the methods described in the previous section, the lack of interpretability of the models is one of the main issues of neural networks.

In particular, the limited understanding of the features learnt by each layer and of the decision process hinder further optimization of the model, its adaptability and transferability to new applications [204]. In fact, it has been shown that neural networks can demonstrate high levels of certainty in their predictions when unrecognizable features are being learnt from the input, which weakens the confidence of experts on these decisions [205].

Understanding neural networks poses a challenge due to the large number of interacting, non-linear parts, as well as the number of learnt parameters. This issue is worsened in deep neural networks due to their size [206]. Without deep and clear understanding of neural networks, the development of better models is reduced to trial and error, which is scientifically unsatisfactory [207].

Deep neural network visualization is an active field of research focused on addressing this interpretability issue. Visualization only started being studied in the 2010s, with the rise of deep learning due to the ImageNet Competition, but significant contributions to the field have already been produced [204]. It is relevant to mention that some of the research groups focused on the development of deep learning models have dedicated themselves to solving this underlying mystery, overlapping the two areas of study.

Current visualization techniques are diverse in what concerns the types of algorithms used, their aim and the information they reveal. While visualization algorithms are becoming progressively diverse, some of them have been more widely used and suffered continuous improvement, becoming increasingly important in the field.

4.5.1 Activation Maximization

Activation maximization was first developed by Erhan et al. as a way to interpret the features learnt by the networks [208]. It attempts to mimic the structure of the visual cortex, building on the hierarchy of the learnt features. To do this, Erhan et al. searched for inputs that maximized the activation of a given unit, based on the idea that a pattern that leads to a strong activation of a neuron may be a good representation of what it is learning. In the first layer, given that the units are linear functions of the input, this method yields particularly clear results.

This search for inputs that maximize the activation can either be done in the training data or it may be transformed in an optimization problem, leading to the synthetic generation of an input that maximizes said activation. This is done through gradient ascent in the input space, changing each pixel of the input in the direction of the gradient in each iteration and keeping the weights constant.

In greater detail, the algorithm can be described as the synthesis of an input x^* that maximized the activation of a given neuron, described as: $x^* = argmax(a_{i,l}(\theta, x))$, where θ are the model parameters. The gradients are computed using backpropagation, with fixed theta values. The input x is updated according to $x \leftarrow x + \eta * \frac{\delta a_{i,l}(\theta,x)}{\delta x}$, where η is the learning rate or step size for the gradient ascent.

Similarly to gradient descent, gradient ascent involves the choice of hyperparameters, namely the learning rate and the number of iterations, which impact the obtained result. The starting input is also a relevant parameter. One can use existing images and change them through gradient ascent, but it is also possible to use random noise or a uniform input.

Figure 4.4 shows the results obtained by Erhan et al. using the MNIST dataset on several layers of the network. It is possible to see that the first layers learn Gabor-like features while subsequent layers learn increasingly complex patterns, looking like pseudo-digits by the third layer. While several random initializations were used, they mostly let to the same input pattern, showing that the activation was unimodal.

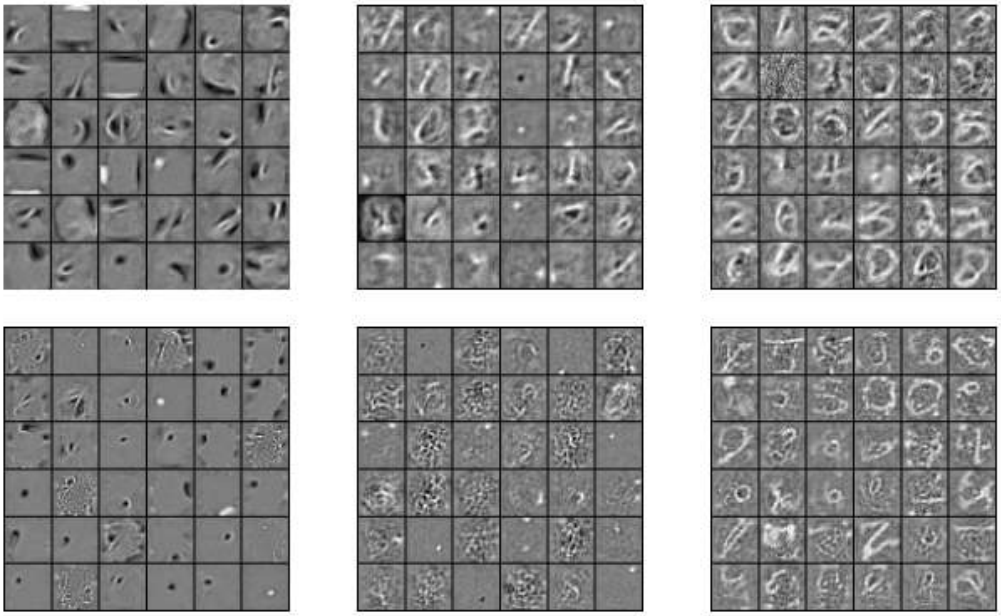


Figure 4.4: Results of the application of activation maximization on the MNIST dataset. The outcome on 36 units is shown, throughout different layers: first layer (left), second layer (center) and third layer (right). Adapted from Erhan et al. [208].

Le et al. [209] used the same principle to verify that their network was learning to identify faces, using images of faces in the test set as well as the previously described optimization method. As Le et al. notes, these methods can be employed complementarily. Using existing images may suffer from fitting to noise while using gradient ascent may lead to a local minimum.

Simonyan et al. [210] built on this, generating an image representative of each class of the dataset used in that work, analogously to the class model of faces produced by Le et al.. Simonyan et al. were also the first to use regularization in the gradient ascent process. In this context, regularization is equivalent to the establishment of image priors, which is relevant when visualizing higher layers, making the patterns more interpretable. This is given by $x^* = argmax(a_{i,l}(\theta,x) - \lambda(x))$, in which $\lambda(x)$ can be any regularization parameter. In this case, the L2 norm was used, preventing pixels with extreme values from dominating the patterns. Other functions can be used for this purpose, including gaussian blur, which penalizes high frequency information [206]. Total variation [211], jitter [212] or data-driven priors [213] may also be used. An alternative to regularization is the use of a generator network to update the input pixels [205]. Generative Adversarial Networks can be used for this purpose.

Google developed a technique called Deep Dreaming, based on activation maximization, using a positive feedback loop to enhance what the network is seeing in a certain image [212]. This overinterpretation is highly dependent on the layer used and the training data. While lower-level layers will represent mostly orientations or other generic features, higher-level layers will showcase the complex, input-like features that it has learnt, depending on the type of images it was trained on. It is possible to use test images as input, but it is also possible to apply this technique to random noise, yielding an output entirely based on the network's knowledge. While the aim of

Deep Dreaming was to get a qualitative sense of the abstraction achieved by a particular layer, it has been used for artistic purposes due to its visually interesting outputs.

4.5.2 Deconvolutional Networks

Deconvolutional networks (DeconvNets) aim to explain what the neural network is doing from the perspective of the input image, by finding the pattern of the input that activates a specific neuron. To do this, the feature map of the neuron is projected to the dimension of the input by means of a neural network that performs the inverse operations of a convolutional neural network – a deconvolutional neural network [204].

DeconvNets were developed by Zeiler et al. [214], initially with the aim of reconstructing natural images through generic features. The same type of algorithm was used for hierarchical image decomposition [215] and later for the visualization of hidden features [207].

A DeconvNet can be seen as the reverse of a convolutional network, as shown in Figure 4.5. It includes deconvolutional layers, which use transposed versions of the filters in the corresponding convolutional layers, as well as unpooling layers, which usually entail the insertion of zeros in the feature maps due to the lack of available information. Finally, reverse rectification layers consist in passing the unpooled feature maps through a ReLu function.

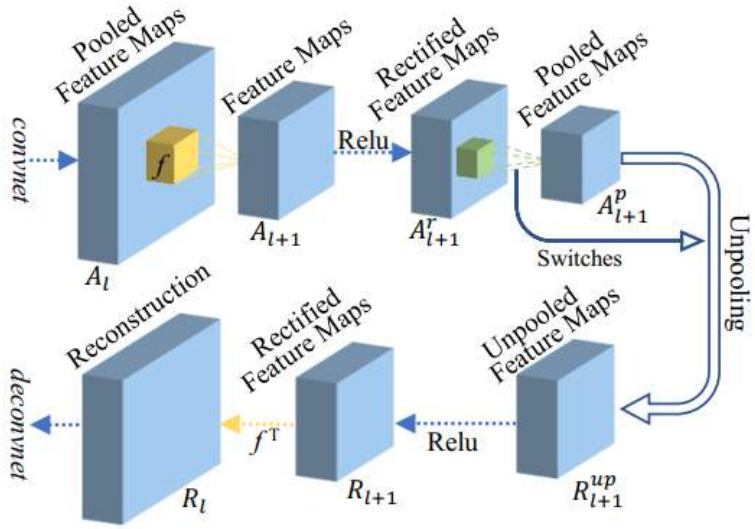


Figure 4.5: Structure of a DeconvNet compared with the original convolutional network [204].

To perform visualization with DeconvNets, the feature maps from all the neurons are captured when an image is used as input. Then, the feature map of the selected neuron is kept, setting all others to zero. The chosen feature map is projected to the dimension of the image using the DeconvNet, and this process is repeated in order to visualize several neurons. Aside from enabling the visualization of the patterns responsible for activation of specific neurons, this technique provides some insight in what concerns the training process of the network. If the DeconvNet finds noisy

patterns, this may indicate that the network has not been trained long enough, or that overfitting is occurring.

4.5.3 Network Inversion

While the previously described techniques aim to visualize the network at a neuronal level, network inversion's goal is to provide a comprehensive perspective of higher structures (i.e. layers or the whole network). With this technique, the feature maps from all the neurons of a layer are used to reconstruct the input. It projects a layer's feature map onto the image dimension, showcasing the features learnt by the layer itself [204].

The idea behind this algorithm was first used in the visualization of traditional computer vision methods, such as Local Binary Descriptors (LBD) [216]. Two adaptations were proposed for visualization in convolutional networks: regularizer-based network inversion [211, 217] and the UpconvNet-based network inversion [218, 219].

The regularizer-based network inversion uses the same network architecture and parameters of the original network. It aims to reconstruct an image, x^*, minimizing the error between the target feature map of the reconstruction and that of the original image. This can be quantified through the expression $x^* = argmax(C * L(A(x) - A(x_0)) - \lambda(x))$ where L is the loss function defined between the feature maps (usually the Euclidean distance) and λ is the regularizer, which restricts x^* to a natural image. This can be an alpha-norm regularizer or it can take on any other form, including the ones discussed in Section 4.5.1. C is a constant that trades off the loss and the regularizer.

To perform this minimization, one starts with an image made of noise (or a hand-made prior), x_0, computing the feature map of its target layer, $A(x_0)$, as well as that of the original image, $A(x)$. Each pixel of the input, x_0, is iteratively changed through gradient descent. This process stops at a stage, x^*, in which the feature map preserves the information retained by the target layer. The main issue with this technique is the computation time, which is relatively large due to the computation of the gradient. The need for regularization to ensure a natural image can also be seen as a disadvantage.

The UpconvNet method has a more complex implementation, since a whole network must be built, but it entails several advantages. The computational effort is made only once, to train the network, with the feature maps being obtained with a single forward pass after this process. Furthermore, the UpconvNet implicitly learns the natural image prior without the need for regularization.

Given a feature vector, the UpconvNet is able to predict the average of the input images that could have led to that vector. To do this, the network includes reversed convolutional layers, as well as reversed rectification and reversed pooling. While the DeconvNet transposes the filters of the original network, the UpconvNet retrains them, based on a set of training images and their feature vectors. It then uses a leaky ReLu function to ensure all the feature maps are positive, as well as a 2-fold upsampling.

Dosovitskiy et al. were able to prove, through the application of the UpconvNet to deep neural networks, that this method led to better visualization quality than the previous one, in particular

in fully connected layers (see Figure 4.6) [219]. Furthermore, the authors showed that all layers preserve the colours and rough position of the objects, and that the encoding of this information varies according to the type of layer: fully connected layers hold information on the probability values, higher convolutional layers on the non-zero activation and lower convolutional layers store it on the weights themselves.

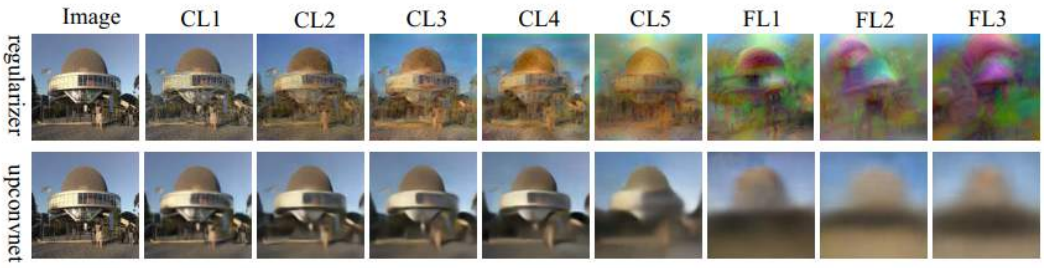


Figure 4.6: Results of the reconstruction of AlexNet through application of regularizer-based network inversion and UpconvNet. [204], adapted from [219].

4.5.4 Network Dissection

Unlike the previously described categories of methods, which aim to reveal the patterns that a neuron or layer is able to capture, network dissection's goal is to relate what is learnt to a semantic concept, linking perceptible visible patterns to easily interpretable concepts [204].

This method was first described by Bau et al. [220], who took the Broadly and Densely named (Broden) dataset and used a single forward pass through each tested network (without any additional training or backpropagation) to assess the correlation between the activation of each unit and each concept. The Broden dataset is comprised of several smaller datasets that are labeled, covering a wide range of object classes, scenes and textures. The algorithm used by Bau et al. started by calculating the activation map for each neuron, given an image from the Broden dataset. For the maps with activations higher than a certain threshold, upsampling with bilinear interpolation was applied in order to restore the dimension of the input. Thresholding was applied to select the regions with high activation, which were compared to the labels in the ground truth using the intersection-over-union score. If this value was high, the neuron that produced the activation map in study was said to be a detector for the concept. Counting the number of concepts aligned with neurons in a layer (unique detectors), one can estimate the interpretability of said layer.

Bau et al.'s method assumes that each concept can be linked to a single neuron. However, it is not illogical to consider that a set of neurons might work in tandem to detect a concept. With this in mind, Fong et al. developed an alternative method, Net2Vec, with the same goal [221]. This algorithm learns a set of weights (using stochastic gradient descent) for each concept to linearly combine the activations of a set of k filters of a given layer. This is used to determine the sets of neurons with higher activation. Similarly to the previous method, the intersection-over-union score is used to determine which sets are detectors of a concept.

With Net2Vec, the authors were able to prove that feature representation is distributed and, by varying the hyperparameter k, it was possible to show that smaller sets led to higher activation (better detection). Other relevant conclusions drawn from network dissection by Bau et al. include the low impact of initialization conditions on interpretability and the decrease in interpretability when batch normalization is used.

4.6 Deep Learning in Health

The application of machine learning to healthcare was vastly motivated by the increase in volume, source diversity and complexity of healthcare data that arose from the developments in areas such as medical imaging, genomics, electronic health records and pervasive sensing, among others. This made traditional analytic methods unable to deal with the rapid influx of data, often unstructured and poorly annotated [222, 223]. Other characteristics of current healthcare data include high dimensionality, heterogeneity, temporal dependency, sparsity and irregularity [223]. Therefore, for a machine learning method to be useful in this area, it must be able to predict and classify with high accuracy despite these characteristics.

Deep learning methods brought a change in the basic assumptions of AI algorithm design, eliminating the need for a supervised definition of the feature space and thus allowing the discovery of novel and more sophisticated features [223, 224]. The ability to handle data multi-modality and perform end-to-end learning are also differentiating factors of these algorithms [223]. While traditional machine learning algorithms have proved to be very useful in some applications, it is expected that the potential that deep learning has shown in areas such as computer vision and natural speech processing can be translated to healthcare, vastly improving the current methods [223, 225].

4.6.1 Current Limitations

While the potential of deep learning in healthcare is undeniable, there are still some limitations that must be taken into account when thinking of applying this type of algorithms.

One of the issues related to deep learning is the lack of interpretability of the models. Methods aiming to be applied in healthcare should be transparent and able to justify its decision, allowing the theoretical and clinical plausibility of the result to be checked [197, 222]. Understanding the results also increases their reliability and trustworthiness by experts. To solve, or at least attenuate, the lack of interpretability of deep learning models, several visualization methods have been developed (see Section 4.5 for more details). However, in some cases, the benefits of a robust and reliable classifier that has a strong correlation with diagnosis of therapy are enough to justify the use of 'black box' models.

The possibility of overfitting and the need for large amounts of data can also be pointed out as limitations of deep learning. While overfitting can occur with any machine learning algorithm, the likelihood of it happening on a deep learning model trained on a small dataset is much higher. Imbalanced datasets such as those involving newly discovered or rare diseases may also lead to

overfitting [224]. Thus, if the amount of data available for a particular problem is small, traditional machine learning methods should be preferred over deep learning approaches. In alternative, overfitting reduction techniques such as those described in Section 4.4.2 may be applied.

The computational cost of training an artificial neural network is sometimes considered an issue of deep learning [226]. However, while training these algorithms can be time-consuming, testing (i.e. classifying new samples or making predictions) with a neural network only involves one forward pass through the network, which usually takes mere seconds. In healthcare applications, if enough data is available to train the algorithm offline, only testing is required on a daily basis. In this case, time should not be considered a limitation for applying a deep learning method.

Finally, choosing the 'right' type of neural network, as well as its architecture, is often seen as a potential problem [226]. This can only be solved with a thorough literature review of similar applications, as well as experimentation. However, this should not be regarded as an issue but as an opportunity for further research and discovery.

Chapter 5

State of The Art - Machine Learning in Epilepsy

Machine learning methods have been widely employed in EEG analysis and in their transversal application across areas of knowledge, from brain-computer interfaces to the diagnosis, monitoring and treatment of diseases (see Section 2.5 for more details). The application of Deep Learning methods for this purpose is now growing, along with the popularity of the field itself, as mentioned in Section 4.1.

The automation of EEG analysis for the purpose of aiding in the diagnosis of epilepsy started in the early 1970s [227]. Automated seizure detection is of the utmost importance, as it aids in diagnosis and can potentially help in the therapeutic process as well. The same can be said for the detection of transient patterns such as interictal discharges, as these can also be helpful for diagnosis. Predicting the occurrence of seizures is equally important, as it can allow timely therapy or pre-emptive actions. In what concerns the treatment of the disease, namely with neurostimulation, the analysis of the EEG can also be of use to optimize the closed loop processes used [228].

The following subsections will focus on the work that has been developed in this area, with more focus being given to the detection of interictal epileptiform discharges, as it will be a cornerstone subject of this dissertation.

5.1 Epileptic Seizure Detection

Systems designed for the detection of epileptic seizures can tell clinicians that seizures are happening and provide them with useful information for their management and aftermath. This is not trivial since it is not common for the patient to have their EEG recorded unless they are already in the hospital due to previous seizures or other complications. The likelihood of capturing EEG signals from ictal periods can be increased through ambulatory EEG recordings or video EEG monitoring [229].

Despite these constrains in data acquisition, there has been extensive work developed in seizure detection. Most algorithms start with the extraction of values, patterns or other biomarkers (i.e.

features) from the EEG and then perform classification of the signal as ictal or normal (binary problem), or as ictal, interictal or normal (multi-class problem) [230].

The first steps in the field were reported by Gotman [231, 232] using mimetic methods. These are rule-based systems that collect features that mimic those used by experts in their classification process, aiming to reproduce it in an automated manner. The use of ANNs with supervised [233] and unsupervised learning methods [234] aimed at reducing the number of parameters considered when assessing if an EEG segment was ictal.

Over time, the automated analysis of EEG signals moved from descriptive and heuristic methods to more advanced approaches, including time and frequency analysis. Wavelet transform and its variants have been some of the most widely used techniques for feature extraction in seizure detection [235–239]. Wavelet transforms decompose the signals into time components at multiple levels of resolution, which can be processed further or used as input of a classifier. Other techniques such as spectral analysis through the computation of the Fourier transform have also been extensively applied in feature extraction [240, 241].

Principal component analysis (PCA) was one of the approaches used by several authors [235, 242] to reduce the dimensionality of the feature space. Subasi et al. [236] compared PCA to other reduction techniques such as independent component analysis (ICA) and linear discriminant analysis (LDA). This study proved that using any of these techniques led to an improvement in the classifier's performance (in this case, an SVM). It also showed that LDA was the most effective method to increase the accuracy of the classifier, but it took considerably longer to train than the other alternatives.

Non-linear methods, such as approximate entropy calculation have also been used for seizure detection [243]. Using approximate entropy analysis, Ocak et al. [239] concluded that normal EEG behaves like a gaussian linear stochastic process while ictal signals showcase a higher degree of nonlinearity, allowing successful detection of these segments. Kannathal et al. [244] extracted several entropy measures from the EEG, namely spectral entropy, Renyi's entropy and Kalmogorov-Sinai entropy, as well as approximate entropy. This work shows that entropies are smaller during seizures, showing a reduction in the intra-cortical information flow, related to an overall decrease in neuronal processing during the ictal stage.

In what concerns the classification process itself, the trend moved from mimicking experts [231, 232] to using more advanced machine learning classifiers such as KNN [245], SVMs [236, 240] and decision trees [241, 246]. Other classifiers such as gaussian mixture models [235] and mixture of experts (a type of algorithm where several learners specialize on certain areas or tasks and a gating network chooses which learner to use in each case) [238] have also been successful in this task. The mixture of experts employed by Subasi et al. [238] consisted in a set of linear classifiers that specialized in different parts of the signal, with the output of the classifier being a mixture of the outputs of the individual classifiers, weighed by their level of expertise in a specific area.

Neural networks of several types have more recently been used as classifiers instead of performing feature extraction [233, 234]. Most approaches so far [237, 242] use extracted features

as inputs for these networks. Jahankhani et al. [237] used statistical information extracted from wavelet coefficients to compare the performance of a multilayer perceptron (MLP) and a radial basis function network (RBF). While the detection accuracy was similar (97% for the MLP and 98% for the RBF), the RBF network was significantly faster during training. The innovative work by Ghosh-Dastidar et al. [247, 248] presented the application of a multi-spiking neural network to the seizure detection task. This algorithm used EEG wavelet features as input and trained the ANN, in which the connection between two neurons was done through multiple synapses.

Newer methods like the one developed by Acharya et al. [229] applied a 13-layer CNN end-to-end, allowing the network to perform both extraction of discriminant features and classification. To train this network, Acharya et al. used a dataset from Bonn University, in which the CNN yielded an accuracy of 88.7%, 95% sensitivity and 90% specificity. This dataset has been used in other papers aiming to detect seizures, enabling comparison of the algorithms' performances. For instance, Ghosh-Dastidar et al.'s multi-spiking neural network reported 92.5% accuracy [247,248], while Martis et al.'s approach using empirical mode decomposition and decision trees led to 95.3% accuracy, 98% sensitivity and 97% specificity [246].

Acharya et al.'s CNN was far from achieving the best reported performance on this dataset, not even being able to beat Acharya et al.'s approach with highest accuracy [249] (99.7%), entailing the extraction of non-linear features and the use of a fuzzy classifier. However, it showed that CNNs can successfully detect epileptic seizures using the raw EEG signal as input, allowing the fusion of the feature extraction and classification steps, getting closer to the current trend in applications of neural networks [136].

5.2 Epileptic Seizure Prediction

Seizure prediction systems must be able to warn clinicians regarding the risk of a patient having a seizure in the near future by detecting pre-ictal changes [230]. Epileptic seizures (usually generalized seizures) are sometimes caused by abrupt transitions, without presenting changes in the EEG in the pre-ictal period [250]. However, many seizures (particularly focal ones) are preceded by several physiological changes such as increased cerebral blood flow and oxygen availability, changes in heart rate, among others. Small groups of neurons (known as bursters) start showing abnormal electrical discharges, aiming to recruit neighbouring neurons [227]. All of these changes translate to alterations in the EEG signal that can be used to predict the imminence of an epileptic seizure.

In 1975, Viglione et al. [251] took the first steps in this field. However, this work, as well as the ones that succeeded it until the early 2000s, focused only on recordings of the pre-ictal period. They did not include interictal recordings, neglecting specificity [227]. From 2003 onwards, several studies on large databases challenged the predictions made in previous works due to their lack of statistical significance and overoptimistic results. Several guidelines and frameworks were proposed to ensure the quality of further work [252], focusing mainly on the prospective proof of the algorithms' predictive power.

While the first studies on seizure prediction used scalp EEGs [251], intracranial EEG (iEEG) has been more widely used for this task since the 1990s [227]. EPILEPSIAE is currently the largest database of EEG and iEEG recordings used for seizure prediction, including data from 275 patients [252].

A plethora of different methods can be used for seizure prediction using normal, pre-ictal and interictal EEG recordings. Similarly to what was previously reported for seizure detection, most approaches to seizure prediction still entail a first step of feature extraction followed by classification by simple thresholding or other machine learning technique [252]. The extracted features can be broadly classified as linear or non-linear and univariate, if they are extracted from a single channel, or multivariate, if extracted from several channels [253].

Mormann et al. [254] compared the performance obtained with features from each of the aforementioned categories in the seizure prediction task. The authors compared linear univariate linear measures (statistical moments, spectral band power, autocorrelation), univariate non-linear measures (estimate of an effective correlation dimension, largest Lyapunov exponent, local flow, algorithmic complexity, surrogate time series and surrogate correction, loss of recurrence), bivariate linear measures (maximum linear cross-correlation) and non-linear measures (non-linear interdependency, phase synchronization). They were able to conclude that univariate measures seemed to be more sensitive to changes right before a seizure while bivariate measures were able to follow dynamic changes up to hours before a seizure. In what concerned linearity, linear measures performed as well as non-linear ones, showing that the presence of non-linearity on the signal itself may not be directly related to the significance of these more complex measures [227]. However, the discriminant power or superiority of a measure or type of measure was not proven, and the authors indicated a combination of univariate and bivariate measures as the most promising path to seizure detection.

Aarabi et al. [255] followed this suggestion and developed an algorithm based on five univariate measures (correlation dimension, correlation entropy, noise level, Lempel-Ziv complexity and largest Lyapunov exponent) and non-linear interdependence, a bivariate measure. Classification was carried out per patient, using a rule-based system that integrated the results of the feature extraction stage. This method yielded a 79.9% sensitivity 30 minutes before the seizure and 90.2% sensitivity 50 minutes before the seizure, at 97% specificity.

Later, Aarabi et al. [256] explored another approach using only univariate features that were used to create patient specific neural mass models to simulate the brain's dynamics. This led to 82.9% sensitivity 30 minutes prior to the seizure and 90.1% sensitivity 50 minutes before the seizure at 100% specificity. While this method showed higher performance than the one using a combination of univariate and bivariate features, the second study was performed on EEGs from a larger number of patients (21 against the 11 used in the previous work [255]), even though these were drawn from the same database (FPSEEG, a subset of EPILEPSIAE). Furthermore, the application of the neural mass model could have had an impact on performance, since the previous work used only the features.

While the previous work focused only on iEEG signals, Bandarabadi et al. [257] took advantage of both types of data present in EPILEPSIAE and compared the performance of an algorithm based on the relative combinations of sub-band spectral powers on scalp EEG and iEEG. The features were selected to reduce dimensionality and fed to an SVM. The authors reported similar performance on both types of data, with iEEG leading to a slightly higher sensitivity and needing less features to reach the same FPR.

Despite the availability of a wide range of classifiers, most methods for seizure prediction use thresholding [255] or SVMs [257]. Mirowski et al. [258] experimented with different classifiers, namely logistic regressors and CNNs. CNNs led to the best performance on 15 patients of the FPSEEG database, yielding 71% sensitivity at 100% specificity, 50 minutes before the seizure. RNNs have also been used as a classifier, using the raw signal and the result of its wavelet decomposition as input [259]. However, this study was done using data from only two patients, lacking significance.

The use of end-to-end ANNs seems like a logical next step in seizure prediction research, as it can extract features that differ from the ones currently being used and lead to a higher performance.

5.3 Treatment Optimization

Currently, epilepsy treatment is done in a scheduled manner or when motivated by the occurrence of a seizure. It may entail the administration of AEDs, VNS or, when it is necessary and possible to identify the origin of the seizure, epilepsy surgery (see Section 3.5 for more details). Closed-loop systems can be used as an alternative way to monitor patients and administer treatment. These systems are based on the continuous monitoring of the patient's signals (in this case, the EEG signal), which can then be used in seizure detection or prediction [230].

The response of a closed-loop system can be warning the patient, a family member of a clinician. Some devices for this purpose are already being commercialized, such as the SmartWatch and Epilert, which send alarms to the smartphones of caretakers when they detect the beginning of a seizure [230]. The system's response can also include real-time therapy administration upon imminence or risk of a seizure. This can stop the occurrence of the seizure altogether or at least alleviate some of its effects, making treatment more efficient [230]. An example of this type of response is NeuroPace, a neurostimulation implantable device that delivers pulses when possible seizure activity is detected instead of providing a stimulus continuously or periodically. This reduces the battery use compared to continuous neurostimulators and reduces the side effects of long-term stimulation [252]. However, the clinician is responsible for the determination of the type of epileptic patterns of the patient and this device can only be used in patients with focal epilepsy with one or two well-defined sites of epileptic activity origin.

As previously discussed, detection algorithms are currently able to classify EEG patterns with high sensitivity and specificity, and thus are good candidates for treatment optimization systems. Prediction algorithms, on the other hand, still present some limitations and, while they show great potential, their applicability to closed-loop therapy delivery is still relatively low [227]. However,

the application of seizure detection algorithms in closed-loop systems still has some limitations, such as the choice of parameters for the detection itself (as seen in the section concerning epileptic seizure detection), as well as for the stimulation or drug delivery [260]. A possible way to overcome this limitation is the use of reinforcement learning algorithms. These can adapt to the patient's brain activity and optimize the treatment strategy automatically. While these methods are very promising, studies have been mostly limited to deep brain stimulation in animals [261, 262].

5.4 Interictal Epileptiform Discharge Detection

The detection of interictal epileptiform discharges (IEDs) differs from the applications discussed in the previous sections, as its usefulness is directed towards the diagnosis process and its efficiency and not towards prevention and treatment (although there is a connection between the number of IEDs and treatment in some epilepsy syndromes, such as absence epilepsy). As seen in Section 3.4.2, IEDs are transient patterns that can be observed in up to 80% of the interictal EEG recordings from epileptic patients, thus being an efficient tool for the diagnosis of the disease [61].

Currently, an expert analyses the EEG signal in 10 to 20 second segments, determining if they include one of these patterns or not. With the disadvantages of the visual analysis techniques in mind (see Section 2.4 for more details), the need for automated detection systems for IEDs becomes evident [61].

This is not a trivial task due to a plethora of factors, starting with the complexity related to the human labeling of IEDs. Textbook definitions of these transients are oversimplified and neurophysiologists have not been able to agree on a precise definition. Furthermore, the morphology of the transients and background rhythms are patient dependent, which makes IED identification more complex. The similarities between IEDs and other transients (such as exaggerated alpha activity and sleep related activities), as well as artefacts, further contribute to the difficulties in labeling [263].

Another hindrance is the type of EEG labeling that is performed by experts, particularly in what concerns signals from epileptic patients. EEG signals are often labeled as control or epileptic, or ictal and interictal for patients, with no identification of epochs with transients. This is enough for seizure detection and prediction algorithms like those described in sections 5.1 and 5.2, but EEG data with epochs labeled as IED or not IED is required for the successful training of supervised learning algorithms for IED detection. Labeling from several experts is advised, since different EEGers often identify different events and an average is needed to establish consensus [61].

Despite these issues, extensive research has been carried out aiming to detect IEDs in EEG signals. Table A in the Appendix summarizes the approaches developed between 1979 and 2018, succinctly describing the methods used.

Until 2000, the vast majority of the algorithms were based on mimetic methods, aiming to emulate the analysis process of an expert. Features such as the relative amplitude of the half-waves, duration and sharpness were widely used by several authors [44, 231, 264–269]. Gotman et

al. started by developing a method based on these features [44] and later adapted it to detect IEDs in different states (active wakefulness, quiet wakefulness, desynchronized EEG, phasic EEG and slow EEG) [264, 265]. Hostetler et al.'s algorithm [267], also based on these features, was able to reach 89% consistency in its predictions, surpassing 83% of the EEGers it was compared to.

These features started to be combined with other methods such as expert systems [268, 270, 271] and template matching [269]. Expert systems aimed to integrate contextual information, either spatial or temporal (or both) into the classification process, while template matching compared each IED candidate pattern to a set of templates, classifying it as an epileptic transient if the similarity is above a certain threshold.

Template matching using raw data has also been extensively used for IED detection [15, 272–280]. While some of the results of this type of method are described qualitatively and the differences in training sets, as well as in the templates themselves, render direct result comparison impossible, several of these algorithms yielded satisfactory results. For instance, Lodder et al. [281] reached a mean sensitivity of 90% with 2.36 false positive detections per minute and Nonclercq et al. [277] reached 90.6% sensitivity at 89.9% selectivity.

Another type of method described in the literature involves the use of mimetic features as input for an ANN, which acts as a classifier [282–285]. Other types of features, such as those derived from wavelet [245, 286–300] and Hilbert transforms [301], as well as Lyapunov exponents [288, 289, 302] have also been used to this end. The use of wavelet transforms in this context was motivated by its ability to perform a multi-resolution analysis in the frequency domain, overcoming the limitations posed by other methods such as the Fourier transform. Wavelet transforms have led to good results, such as those reported by Artameeyanant et al. [296] (76.55% sensitivity, 81.30% specificity, 89.47% accuracy) and Song et al. [300] (96.0% sensitivity, 93.6% specificity, 94.8% accuracy).

ANNs have also been combined with expert systems for this task [286, 303, 304]. In this type of methods, ANNs are usually used to pre-classify EEG segments into categories (eg. spikes, muscle activity, eye blinks or sharp alpha activity in Tzallas et al.'s work [304]), which are then classified by the expert system using contextual information. Tzallas et al. achieved 84.44% accuracy using this system, while Argoud et al.'s [303] algorithm yielded 70.78% sensitivity at 69.12% specificity in spike detection and 71.91% sensitivity at 79.19% specificity for sharp waves.

Pang et al. [305] tested several ANN-based methods with different features to compare them and assess their performance on a different dataset than the one that was initially used. The algorithms developed by Webber et al. [282], Kalacyi et al. [306], Ozdamar et al. [307] and Tarassenko et al. [308], based on 3-layer ANNs with varying number of nodes on the input and hidden layers and different input features, were applied to a dataset comprised of records from 7 epilepsy patients and 8 normal controls. Webber et al.'s method, based on mimetic features, led to the best performance, achieving 86.61% sensitivity and 86.32% selectivity.

While many authors have used multi-layer perceptrons (MLPs) or slight variants to detect IEDs, different types of ANNs have been tested for this task. James et al. [283] and Kurth et al. [309] used a Kohonen's self-organizing feature map, which is a single layer ANN with neurons that

get specifically tuned for a certain input pattern through unsupervised learning. The LAMSTAR network, which uses self-organizing maps combined with statistical decision tools, was employed by Nigam et al. [310]. Other feedforward network approaches included Wilson et al.'s monotonic network [284], Ubeyli et al.'s [311] mixture of experts and Song et al.'s [300] extreme learning machine, which transforms the learning problem into a linear system through which the weights can be determined. Ubeyli et al. used probabilistic neural networks, which handle multi-class problems by decomposing it into dichotomies that can be decided by neurons, in several works [312, 313]. Aside from feed-forward networks, RNNs have also been an option for IED detection [288, 291, 297, 314]. The use of CNNs as an end-to-end classifier [315, 316] is starting to grow, as these networks have shown potential in many different applications. CNNs have even been used as a feature extractor by Thomas et al. [317], followed by classification with an SVM. This algorithm yielded 83.86% accuracy and 55% precision at 80% sensitivity. Machine learning classifiers such as SVMs have been extensively for this task [289, 290, 318–321] and others such as KNN [245, 294, 322] and genetic algorithms [245, 293] have also been employed.

The vast majority of these studies were developed using proprietary datasets, hindering the possibility of result comparison. However, a five-class dataset was made public by Andrzejak [323], with each set including 100 single channel EEG segments of 23.6s without discernible artefacts. Sets A and B (intracranial) were recorded from 5 normal controls with eyes open and closed, respectively; sets C and D (extracranial) were recorded from 5 epilepsy patients in seizure free intervals in the epileptogenetic zone (D) and hippocampal formation of the opposite side of the brain (C); set E contained seizure activity from those patients. Several authors have used this dataset or parts of it [245, 288–292, 294, 295, 302, 311–313, 322, 324–326], enabling some further comparisons. Guler et al. achieved 68.8% accuracy with a MLP, 72.0% with a PNN, 75.6% with an SVM, using wavelet transform and Lyapunov exponents [288] and 96.79% accuracy with an RNN, using Lyapunov exponents as input [324]. Ubeyli et al.'s PNN trained with Lyapunov exponents as input [313] yielded 98.05% accuracy, showing that Lyapunov exponents may be more appropriate than their combination with wavelet transforms for this purpose. Even higher performance was reached using eigenvector methods for feature extraction, with a PNN yielding 97.63% accuracy [312]. With the same feature extraction methods, the authors obtained 99.3% accuracy with an SVM [290] and 98.15% accuracy with a RNN [291].

Other methods trained using Andrzejak's dataset include the one developed by Guo et al. [245], which combined wavelet transform, genetic algorithms and KNN. It was found that using KNN without the genetic algorithm led to $67.2 \pm 1.2\%$ accuracy, while the combination with the genetic algorithm increased this value to $93.5 \pm 1.2\%$. Orhan et al. [295] tried another approach using the wavelet transform, k-means and a MLP (using the output of k-means as input). This led to 98.80% accuracy, 99.33% specificity and 98.02% sensitivity in the diagnosis of epilepsy. Iscan et al. [326] compared several classifiers, namely SVM, least-squares SVM (LV-SVM), KNN, Parzen window, LDA, decision tree, Naïve Bayes, nearest mean and quadratic classifier. The authors used cross correlation to extract time features and power spectral density to extract frequency features, which were used as input to the classifiers. It was found that the combination of both types of features

increased the accuracy of the algorithms, and that LV-SVM, binary decision trees and quadratic classifiers yielded the highest scores.

Another relevant comparison can be drawn between ANN performance with parameterized and raw data. Webber et al. [282] used a 3-layer network with fully connected layers, with either raw data or mimetic features as input. This algorithm led to an intersection between sensitivity and selectivity of 73% for parameterized data and 43% for raw data. Ko et al. [327] tried to detect IEDs using a MLP and raw EEG data, but the performance of the algorithm was below random, leading the authors to conclude it was impossible to perform this classification on raw data.

This has since been disproven, as CNNs have been shown to detect IEDs with high accuracy. Johansen et al. [316] achieved an AUC value of 0.947 with a CNN comprised of 3 convolutional layers, trained on raw data. Tjepkema et al. [315] trained one and two-dimensional CNNs, as well as LSTMs with raw data. Two-dimensional CNNs yielded 0.94 AUC for the test set, as well as 47.4% sensitivity at 98.0% specificity with only 0.6 false detections per minute. As previously mentioned, the use of CNNs is growing in this field, and training on raw data to achieve end-to-end classifiers has proven to be possible, achieving satisfactory results.

While some algorithms, like Wilson et al.'s SpikeDetector [284], have become commercially available, it is relevant to discuss why the algorithms described in this section, some of which showcasing satisfactory results, are not widely used in clinics.

Some studies present algorithms trained on data from very small groups of patients ([304,328] - 1 patient, [273, 308] - 2 patients). Since IEDs have patient dependent characteristics in what concerns their morphology, a robust algorithm ready for clinical use would have to be trained on a larger database to account for patient variability. Also, many algorithms are trained using routine EEGs, which are shorter than the long-term recordings used for diagnosis and thus contain less artefacts [329]. This reduces the robustness of the algorithms in what concerns artefact occurrence and lowers their potential for clinical use.

Another issue is the lack of a baseline comparison for the algorithms, which is related to the diversity of training datasets (since there is no database used by all authors to either train or test their algorithms) and to the way performance is assessed. Several methods for statistical analysis are used in the studies and the shown metrics are not always comparable. While many report sensitivity and specificity, as well as false positive detections per hour, many studies do not present a detection rate per hour, which has a direct impact on the significance of the results. For instance, if there is a high number of spikes (i.e. high frequency of IED occurrence), it is possible to obtain an algorithm with high sensitivity and specificity, but also with a high rate of false detections [61].

For an algorithm to be used in a clinical context, it must be able to reduce the time spent in signal analysis without compromising performance (preferably improving it). This implies that the computational burden of the algorithm must be low enough to allow online classification after training to avoid long computation times. It also means that training must be previously done, on a large and diverse dataset, to ensure robustness. Algorithms with low sensitivity cannot be used since the expert still needs to spend time looking for the IEDs that were not detected. On the other hand, algorithms with low specificity cannot be used either, due to the number of false detections

that have to be manually rejected [329]. Therefore, a detection rate that is consistent with the frequency of IED occurrences, along with high sensitivity and specificity are crucial requirements for these algorithms.

They must also be user-friendly and present results in a clear and easily interpretable manner to fit in the diagnosis workflow. As most algorithms developed under an academic setting do not include a user interface, since it is not needed in an early stage of development, this means that a large improvement is needed in what concerns user interface and experience, as well as result visualization. Before reaching the clinics, beta testing with clinicians should be carried out in order to fix issues related to these aspects.

With this in mind, it is possible to conclude that the wide majority of the described algorithms could not be used in clinics. It is also worth mentioning that efforts towards the creation of robust, user-friendly software in an academic setting promote technology transfer and may speed up the process of getting an algorithm to the clinics.

Chapter 6

Methods

6.1 EEG data and pre-processing

6.1.1 EEG Data

We used EEG data from 217 patients between 4 and 72 years of age, randomly selected from the digital database of the Medisch Spectrum Twente in the Netherlands. This dataset included interictal EEGs from patients with focal (50 patients) and generalized (49 patients) epilepsy, containing interictal epileptiform discharges (sharp waves, spikes, spikeslow-waves or polyspike-waves: IEDs). We also included EEGs with non-epileptiform abnormalities (51 patients) and normal EEGs (67 patients).

This was done based on the diagnosis and notes from the EEGers (eg. searching for 'focal epilepsy' or 'normal EEG' in the conclusion of the report. The complete clinical report and the EEG recording itself were reviewed by the expert (Michel van Putten or Marleen Tjepkema-Cloostermans). Epileptic EEGs were annotated by the expert so that IEDs could be easily identified. Epochs in which there was uncertainty regarding the occurrence of an IED were not labeled, ensuring that all the annotations corresponded to the unequivocal presence of an epileptiform discharge. In turn, this led to some epochs in epileptic EEGs including unidentified IEDs. Epochs of the Normal and Abnormal classes were not labeled.

The recordings were made with twenty-one silver/silver chloride cup electrodes placed on the scalp according to the international 10-20 system [37]. All EEGs were obtained as part of routine care, and anonymized before further analysis.

6.1.2 EEG pre-processing

EEG data was filtered in the 0.5-35Hz range to reduce artefacts. We downsampled it to 125Hz to reduce input size (and consequently computational complexity). Subsequently, signals were re-referenced to a longitudinal bipolar montage. The 18 channels of this montage are represented in Fig. 6.1, which shows the connection between the numerical order of the channels and their positions in space. While this is not the only montage used in visual analysis when experts are

searching for IEDs (switching between montages is often a useful tool), it is a good approximation of a real-life scenario. We split each recording into 2s non-overlapping epochs, yielding a 18x250 (channels x time) matrix for each epoch. The duration of the epochs was chosen based on the fact that IEDs have less than 1s duration and 2s allows the preservation of temporal context.

The pre-processing routine is summarized in Fig. 6.2 and it was implemented in Matlab R2019a (The MathWorks, Inc., Natick, MA). The resulting epochs were used as input for the neural networks described in 6.2.

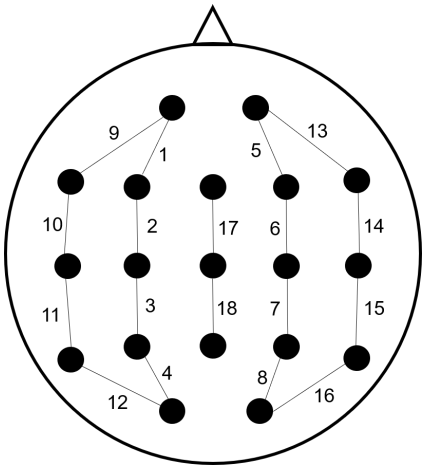


Figure 6.1: Representation of the 18 channels in the longitudinal bipolar montage. Electrodes are represented as black circles, with each channel shown as a connecting line between electrodes.

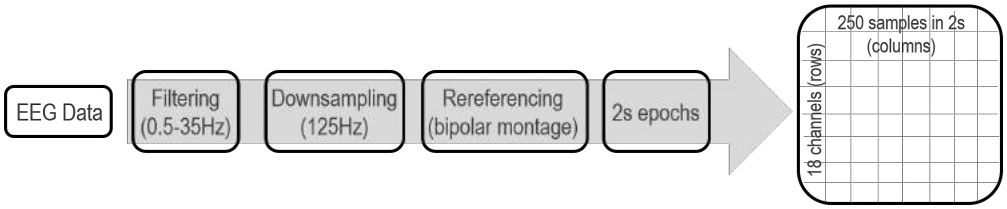


Figure 6.2: Schematic representation of the pre-processing steps applied to all EEG data.

6.1.3 Problem Definition

We tackled several classification problems in this project, and different datasets were created accordingly [1]. The data included in each of these sets, as well as the duration, total number of epochs and number of epochs of the positive class in the corresponding training and test set is shown in Table 6.1.

The first goal was IED detection using Focal and Generalized epilepsy data as positive (label '1') and normal EEG data the negative class ('0'). Two different sets were created for this purpose. Set **A** (see Fig. B.1) included the full extent of the recordings from all the epilepsy patients (both focal and generalized), as well as all the normal recordings. IED epochs from the epileptic

[1] The rationale for the creation of some of the datasets will be discussed further in Chapter 8

Table 6.1: Number of patients, duration, total epochs and epochs of the positive class of each created dataset for the training and test sets. For multi-class problems (sets G through J), the number of epochs of all classes is stated, since there is no positive class.

			Train			Test		
	Classes	Patients	Duration (h)	Epochs	Positive	Duration (h)	Epochs	Positive
Set A	N + F + G	166	62.6	112747	1977	13.2	23774	734
Set B	N + F + G	166	24.3	43867	2220	8.8	15886	452
Set C	N + A + F + G	217	40.5	72279	2015	11.4	20470	658
Set D	F + G	99	1.1	1972	556	0.4	693	294
Set E	N + A	118	39.3	70746	3864	11.1	19916	7738
Set F	N + A + F + G	217	41.3	74330	44525	10.6	19108	12267
Set G	N + F + G	166	50.5	90954	23533 F 20080 G 47341 N	12.4	22354	5695 F 7545 G 8844 N
Set H	N + F + G	166	24.1	43460	1506 F 772 G 41182 N	8.2	14674	309 F 132 G 14233 N
Set I	N + A + F + G	217	64.6	116265	24258 F 21614 G 43306 N 27087 A	17.8	32066	5240 F 6542 G 13460 N 6824 A
Set J	N + A + F + G	217	41.4	74498	1548 F 744 G 44951 N 27275 A	10.5	18854	267 F 109 G 11842 N 6636 A

recordings constituted the positive class, with all the remaining data belonging to the negative class. Set **B** (see Fig. B.2) had the same positive class, but the negative class was only comprised of normal EEGs, with the normal part of the recordings of epilepsy patients being discarded.

Another approach to this problem was the inclusion of EEGs with non-epileptiform abnormalities in the negative class. A new set (set **C**, see Fig. B.3) was built with the same positive class as sets A and B and a negative class including normal and non-epileptic abnormal EEGs (once more, the normal part of the recordings of epilepsy patients was discarded).

We also tackled the problem of distinguishing between Focal and Generalized epilepsy based on IEDs. For that purpose, we created set **D** (see Fig. B.4) containing Focal IEDs as the positive class and Generalized IEDs as the negative class.

Distinguishing normal from abnormal EEGs was also a problem addressed in this project. Set **E** (see Fig. B.5), with normal EEGs as the negative class and abnormal, non-epileptic EEGs labeled '1'. Set **F** (see Fig. B.6) had the same negative class, with the positive one including abnormal EEGs with epileptiform and non-epileptiform abnormalities (only IED epochs from epilepsy patients were used, discarding the normal part of the signal).

We also addressed multi-class problems in this project. The first one concerned the distinction between Focal epilepsy (labeled as '0'), Generalized epilepsy ('1') and normal signals ('2'), while the second one added the abnormal, non-epileptic EEGs as a fourth class ('3'). Set **G** (see Fig. B.7) was created for the 3-class problem, with IEDs from Focal and Generalized epilepsy labeled as '0' and '1', respectively, and the normal part of the signal labeled as '2', having the same label as the EEGs from normal controls. Set **H** (see Fig. B.8) was the same, without the normal part of the EEGs from epilepsy patients. Sets **I** and **J** (see Figs. B.9 and B.10) were the equivalents for the 4-class problem.

6.1.4 Dataset Creation

EEG data was randomized and split into a training/validation set containing 80% of the recordings and a test set comprised of the remaining 20%. All epochs from a patient were used either for training or testing. We applied five-fold cross validation on the training/validation set, further partitioning it in each iteration. One of these partitions was used to validate the model and the others were used for training, changing the validation partition in each iteration. Fig. B.11 in the Appendix illustrates this division.

6.2 Deep Learning Models

The models were implemented in Python 2.7 using Keras 2, Theano and a CUDA-enabled NVIDIA GPU (GTX-1080), running on CentOS 7. Stochastic optimization was performed using an Adam optimizer [191] with a learning rate of $2 * 10^{-5}$, $\beta 1$=0.91, $\beta 2$=0.999, and $\varepsilon = 10^{-8}$. A sparse categorical cross entropy function was used to estimate loss and a batch size of 64 was used.

6.2.1 VGG

The VGG network was created in 2014 in Oxford by Karen Simonyan and Andrew Zisserman from the Visual Geometry Group [176], and it is comprised of several sets of padding and convolutional layers followed by max pooling. After these, flattening precedes a set of two fully connected layers with ReLu activation functions intercalated by dropout. Finally, a fully connected layer with a Softmax activation function produces the output. This simplified architecture can be seen in Fig. 6.3. Fig. B.12 in the Appendix shows the full architecture of the model.

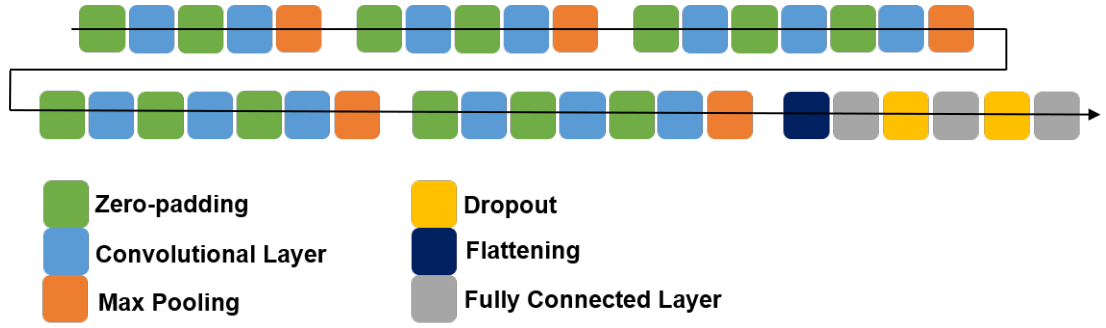


Figure 6.3: Simplified architecture of the VGG C model. Five blocks comprised of padding, convolutional and pooling layers are followed by flattening and three fully connected layers, intercalated with dropout.

The first two sets of layers were comprised of two padding layers intercalated with convolutional layers with 3x3 filters (64 filters in the first set and 128 in the second one), ending with 2x2 max pooling. The following three sets included one more padding and convolutional layer, with 1x1 filters, before pooling. The number of filters doubled to 256 in the convolutional layers of the third set and then again to 512 in the fourth set, which was also the same number of filters used

for the final set. The filters of the last convolutional layer of each block had stride of 2, which changes the receptive field of the neurons and makes filtering faster. The last fully connected layer was changed to 2, 3 or 4 units instead of the original 1000, depending on the problem at hand.

6.2.2 ResNet

The ResNet was created by Google in 2015 [178]. Its depth and the use of residual modules were the main innovations of this model (see Section 4.1.4 for more details), which, in turn, led to an improvement in performance in the ImageNet competition and decrease in computation time. The ResNet is comprised of a first convolutional layer followed by a max pooling layer. Four residual modules with blocks of 2 or 3 layers (depending on the architecture) intercalated with residual connections are repeated several times. Finally, average pooling is followed by flattening and by a fully connected layer with Softmax activation.

While a 152-layer architecture of the ResNet was used in the ImageNet competition, there are variants of this model with less layers, including the ResNet50, which was used in this project. The simplified architecture of the ResNet50 can be seen in Fig. 6.4, with each block representing a residual module as described in 4.1.4. Fig. B.13 in the Appendix shows the full architecture of the model.

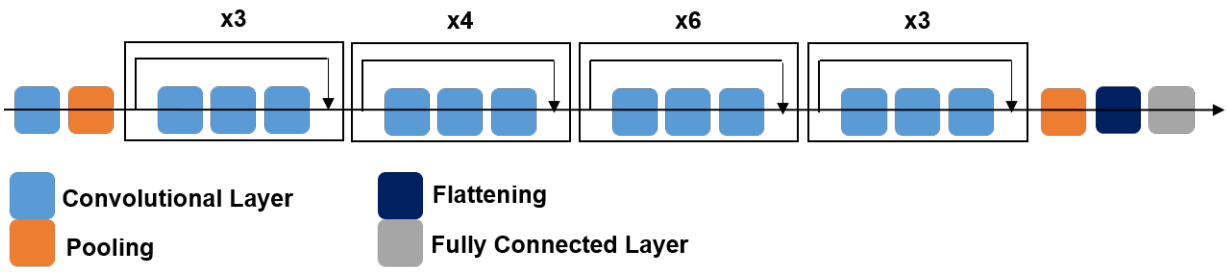


Figure 6.4: Simplified architecture of the ResNet50 model. A convolutional and a pooling layer are followed by four sets of repeated residual modules. Pooling and flattening precede a fully connected layer.

The first layers of the model used in this project were common to all the ResNet architectures. The convolutional layer was comprised of 64 7x7 filters with stride 2, followed by 3x3 max pooling. The residual modules included three convolutional layers with 1x1, 3x3 and 1x1 filters, respectively. The number of filters doubled per module, starting with 64, 64 and 256 in the first one (i.e. the last module had 512, 512 and 2048 filters). The first and last modules were repeated 3 times, the second module was repeated 4 times and the third module 6 times. Finally, average pooling, flattening and a fully connected layer with Softmax activation completed the model. In this project, 2 units were used instead of the original 1000 in the output of the ResNet.

6.2.3 Custom-made models

We used two other models in this project, which were not based on any networks from the literature. The following sections describe the architectures of these models.

6.2.3.1 M1

The first model, named M1, was comprised of a 1x1 zero-padding layer, followed by a convolutional layer with 128 3x3 filters and ReLu activation. Max pooling with 2x2 pool size and strides ensued, followed by a 0.5 dropout. Flattening and a fully connected layer with Softmax activation and 2 output units completed the model. Fig. 6.5 shows the simplified architecture of this model (see Appendix Fig. B.14 for the detailed architecture).

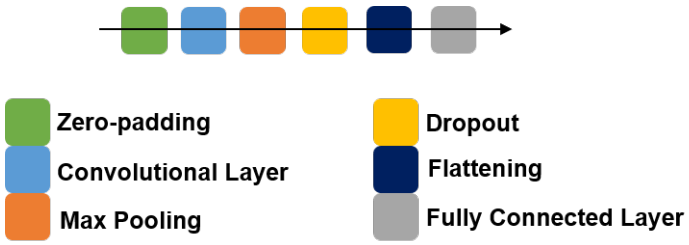


Figure 6.5: Simplified architecture of the M1 model. The first layer performs zero-padding, followed by a convolutional, pooling and dropout layers. Flattening precedes the fully connected layer.

6.2.3.2 M2

The second model, named M2, included the same initial three layers of M1. These were followed by another set of a convolutional layer, max pooling and dropout with the same parameters. Flattening preceded three fully connected layers. The first two had ReLu activation, with 128 and 64 output units, respectively. The final fully connected layer had Softmax activation and two units. Fig. 6.6 shows the simplified architecture of this model (see Appendix Fig. B.15 for the detailed architecture).

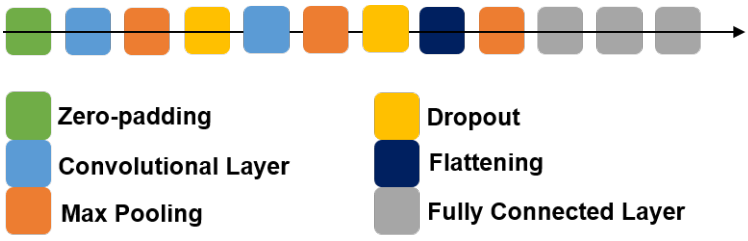


Figure 6.6: Simplified architecture of the M2 model. The first layer performs zero-padding. Two blocks of convolutional, pooling and dropout layers follow. Flattening precedes three fully connected layers.

6.3 Visualization Techniques

6.3.1 Filter Visualization

To visualize the filters, we calculated and normalized the gradients of the input with regard to the loss for each filter of each convolutional layer. Then, starting with a synthetically generated gray image with random noise with the same dimensions as the EEG epochs, we performed gradient ascent for 10 thousand iterations with a step size of 0.1. The filters with the highest loss values in the end were assumed to have a more relevant shape and, as such, only the best 64 filters according to this criterion were plotted and analysed.

6.3.2 Input Maximization

For this visualization technique, instead of staring with random noise, the inputs of the dataset itself were used. The border points of the input were not included in the analysis to avoid border artefacts. Similarly to what was done in filter visualization, we calculated and normalized the gradient and performed gradient ascent for 10 thousand iterations and a step size of 0.2. The areas of the input where the value at the end of this process was higher were the ones that had led to the highest activation of the filters and were plotted with warmer colors. The rationale for this technique is based on the idea that a high response to a certain pattern could be a good initial representation of what the hidden unit is doing [208].

6.3.3 Occlusion

This technique aims to show the areas of the input of highest importance for classification by covering patches of the input iteratively and calculating the network's response to the resulting signal. In this case, a significant change in the response shows that the covered patch at hand had a great impact on the way the model made its decision [207].

To do this, a grid was applied to each sample and, in each iteration, the contents of a patch of that grid were set to zero, leaving the remaining sample untouched. The probability resulting from the network's prediction was stored in the center of the occluded patch so that it could be compared to the prediction without occlusion. After going through the whole image, we calculated the difference between the value in each grid patch center and the original prediction of the network, and the patches with higher differences were plotted with warmer colors. The dimensions of the grid varied between 10 and 50 for the time axis and between 1 and 6 for the channel axis.

6.4 Performance assessment

6.4.1 Binary problems

Receiver Operating Characteristic (ROC) curves were calculated for each of the cross-validation iterations using 101 discretizations. This was then averaged, yielding an average ROC curve for

each set. The area under said curve (AUC) was calculated. Confidence Intervals (CIs) at 95% were calculated for the ROCs and AUCs.

We calculated the Sensitivity, Specificity, the number of True Positives, True Negatives, False Positives and False Negatives, as well as the corresponding rates per hour using a threshold of 0.5 and at another threshold where the values of Sensitivity and Specificity were as similar as possible (achieved by calculating the minimum of the difference). Confidence Intervals at 95% were calculated for these parameters. The False Positive rate per hour was also assessed at a Specificity of 99%. These routines were implemented in Matlab R2019a (The MathWorks, Inc., Natick, MA).

Confusion matrices using a threshold of 0.5 were generated in Python2.7.

6.4.2 Multi-class problems

We generated two types of ROC curves for multi-class problems. The first type treated the problem as binary (one-against-all), choosing one of the classes as positive and the remaining as negative. This yielded a set of ROC curves, one for each class, obtained using the same method as described above for binary problems. This was implemented in Matlab R2019a (The MathWorks, Inc., Natick, MA).

The second type of ROC curves took a more global approach, combining all the classes using macro-averaging, creating a single ROC curve per problem. Macro-averaging was used as it calculated the metrics for each class independently, weighing them equally, which reduced the potential impact of class imbalance in performance assessment (since micro-averaging would have weighed the classes based on their relative size, giving more importance to correct classifications in classes containing more samples). The routine used to generate these curves was implemented in Python2.7.

The Accuracy, Sensitivity and Specificity were calculated per class using a threshold where the values of Sensitivity and Specificity were as similar as possible, analogously to what was described for binary problems. This routine was also implemented in Matlab R2019a (The MathWorks, Inc., Natick, MA).

Chapter 7

Results

7.1 IED detection

The main goal of this thesis was the detection of IEDs (i.e. obtaining the probability of an EEG epoch containing an IED) using the models described in 6.2. The four networks were trained for this purpose. The results yielded by the models for this classification problem are shown below.

7.1.1 Normal vs IEDs with full epileptic EEG - Set A

We first trained the VGG with the full EEG recordings of epilepsy patients (containing normal epochs and IEDs) and normal controls, weighing both classes equally. When applied to the test set, this approach yielded the confusion matrix at a threshold of 0.5 (and respective normalization) shown in Fig. C.1. This corresponded to 97.0% accuracy, with 23070 out of 23774 samples being classified correctly. However, 82.6% of IEDs were misclassified as normal.

Different weights were then assigned to both classes, with the positive class (i.e. the IEDs) being weighed more heavily. Fig. C.2 shows the normalized confusion matrices obtained with a threshold of 0.5 when weights of 10, 50 and 100 were assigned to the positive class. Accuracy values at this threshold were 95.1%, 94.2% and 91.5%, with 20.7%, 20.2% and 10.5% of IED epochs being classified as Normal.

All four models were trained using weights 100:1. Fig. C.3 shows the average ROC curves for the training and test sets after 5-fold cross-validation for the different models, with the 95% confidence interval as a shaded area. The AUC values obtained with the VGG model were 0.99 (CI=0.99-0.99) for the training set and 0.91 (CI=0.89-0.94) for the test set. The ResNet yielded 0.94 (CI=0.85-1.00) on the training set and 0.76 (CI=0.69-0.84) on the test set. The AUC values for M1 and M2 were 0.91 (CI=0.88-0.95) and 0.99 (CI=0.99-0.99) for the training set, with 0.70 (CI=0.62-0.78) and 0.91 (CI=0.89-0.94) having been obtained for the test set.

Visualization techniques were applied to the VGG model trained using set A and weights 100:1. Fig. 7.1 show examples of the results yielded by the filter visualization technique. These show the native shape of the filters, since random noise was applied in the input and gradient ascent was performed to maximize loss, aiming to show how the filter looks the first time it is convolved

with the EEG input, before gradient descent (since, then, the filters become different in each pass for the different inputs). These figures can be understood as images with the same dimensions as the 2 second, 18-channel EEG epochs used as input, since those dimensions were also applied to the synthetic input used in the algorithm.

Examples of the results of input maximization are shown in Fig. 7.2. For each epoch, the average activation value over all the filters of all the convolutional layers is shown. The numerical values of the average activation correspond to a color scale where warmer colors equal higher activation, and, as such, areas plotted with warmer colors correspond to higher average activation.

Fig. 7.3 shows examples of the results obtained when occlusion is applied to the VGG network trained with weights 100:1, with set A. Two examples are provided for cases of true positive, false positive, true negative and false negative, along with the probabilities assigned by the model to each epoch. The scale shows the difference between the probability for each epoch and the value obtained with an occluded patch. Higher differences are plotted in warmer colors, showing that removing the patch centered in that area led to a significant change in classification, indicating that that area is important to the networks decision process.

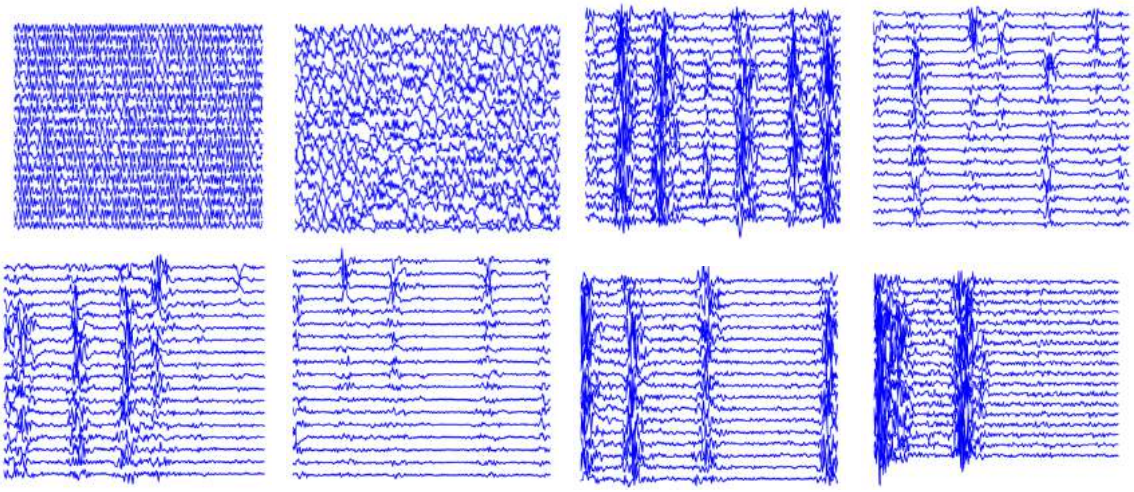


Figure 7.1: Examples of results of the application of filter visualization to the VGG model, trained using set A and weights 100:1. The two first panels (first row, on the left) show filters from lower-level layers; the remaining panels concern filters of higher-level layers. These have the same dimensions as the 2s, 18-channel EEG epochs used as input. They show the native shape of each represented filter, before any forward pass of the network.

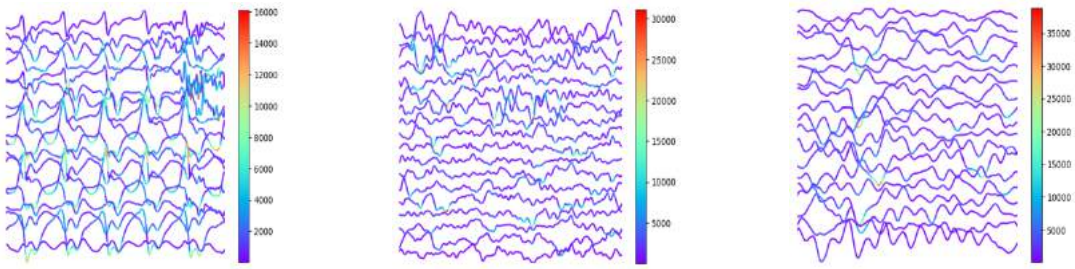


Figure 7.2: Examples of input maximization for the VGG model, trained using set A and weights 100:1. These illustrate three different EEG epochs and the average relative activation caused by each part of the signal over all the filters in the three convolutional layers. Higher activations are plotted in warmer colors and the corresponding numerical values are the average activations.

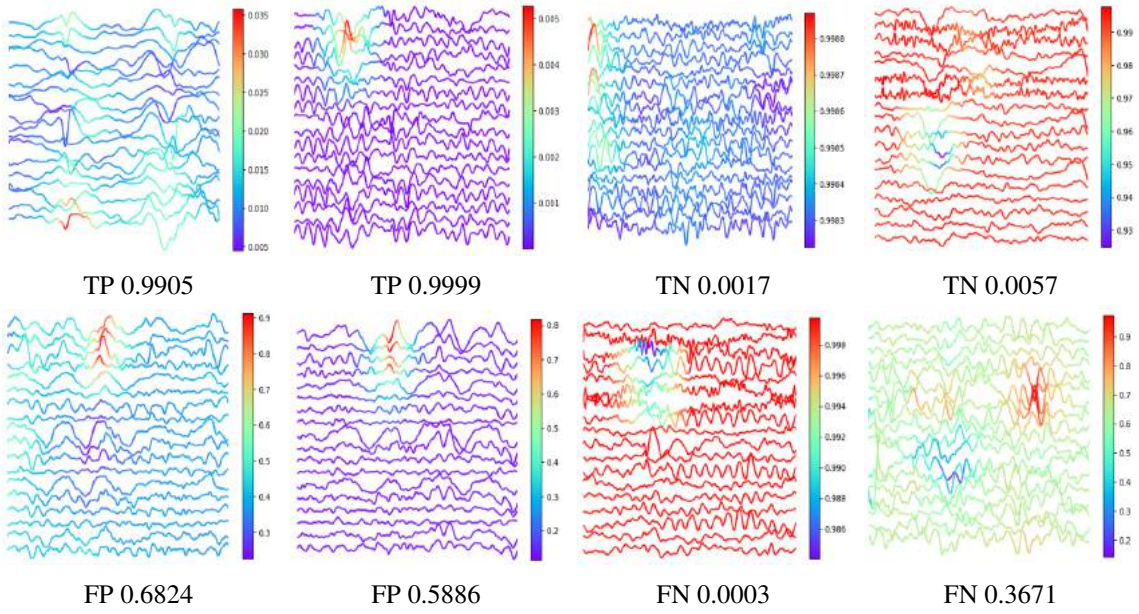


Figure 7.3: Examples of occlusion for the VGG model trained using set A with weights 100:1 in cases of: First row, left: True Positive; first row, right: True Negative; second row, left: False Positive; second row, right: False Negative. The probability assigned by the network regarding the occurrence of an IED in each sample is shown. The scale shows the difference between the probability assigned to the epoch and the probability obtained when a patch is occluded, and warmer colors are assigned to higher differences. Thus, areas plotted in warmer colors are more important for classification.

7.1.2 Normal vs IEDs after removal of normal epochs - Set B

After the removal of all normal epochs from the EEG recordings of epilepsy patients, the training set became about 1/3 of the original training set. All the models were trained using 100:1 weights for the positive class. Fig. 7.4 shows the average ROC curves for the training and test sets after 5-fold cross-validation for the different models, with the 95% confidence interval as a shaded area. The AUC values obtained with the VGG model were 0.99 (CI=0.99-1.00) for the

training set and 0.96 (CI=0.95-0.97) for the test set. The ResNet yielded 0.97 (CI=0.93-1.00) on the training set and 0.89 (CI=0.84-0.94) on the test set. The AUC values for M1 and M2 were 0.94 (CI=0.93-0.96) and 0.96 (CI=0.94-0.98) for the training set, with 0.84 (CI=0.83-0.85) and 0.91 (CI=0.88-0.94) having been obtained for the test set.

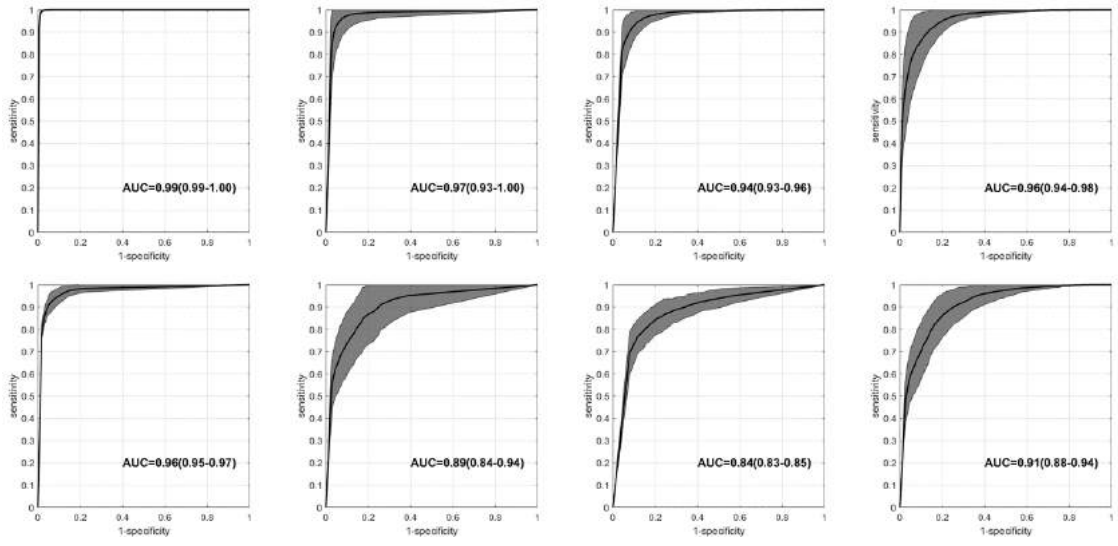
Table 7.1 shows the false positive and true positive rates per hour for the four models at a threshold where the sensitivity equals the specificity. For the VGG, this occurred at 93% sensitivity/specificity for the test set and 98% for the training set. The model led to a false positive rate per hour of 122.41 (CI=27.63-217.20) on the test set and 22.30 (CI=6.32-38.28) on the training set. True positive detections were 47.72 (CI=45.60-49.84) and 91.03 (CI=80.35-100.22) per hour, on the test and training set, respectively. Setting the specificity threshold at 99%, the false positive rates per hour of this model were 16.50 (CI=4.95-28.06) and 32.31 (CI=15.15-49.46), for the training and test set. At 94% sensitivity/specificity, on the training set, the ResNet led to 99.03 (CI=17.96-169.71) false positive detections per hour and 86.41 (74.04-95.87) true positive detections per hour. On the test set, the ResNet achieved 293.79 (CI=79.69-507.89) false positives per hour and 42.98 (CI=35.43-50.37) true positives per hour at a sensitivity/specificity of 83%. Setting the specificity threshold at 99%, the false positive rates per hour of this model were 0.00 (CI=0.00-0.02) and 17.49 (CI=7.05-27-95), for the training and test set. For model M1, false positives per hour were 145.94 (CI=87.30-204.59) and 307.11 (CI=203.93-410.39) on the training and test set, respectively. The true positive rate was 83.55 (CI=70.79-96.31) and 42.20 (CI=38.48-45.93) per hour, at a sensitivity/specificity of 91% for the training set and 82% for the test set. For model M2, this intersection occurred at 88% for the training set and 83% for the test set. false positive detections were 195.02 (CI=7.19-382.85) and 301.85 (CI=76.55-527.14), respectively. The model detected 80.99 (CI=64.47-97.52) true positives per hour on the training set and 42.61 (CI=37.13-48.10) on the test set.

Table C.1 shows the number of epochs, the number of IEDs, the sensitivity and specificity values at a threshold of 0.5, as well as the number of true positives, true negatives, false positives and false negatives at this threshold. The Normal class was classified with an average specificity of 98.64%, with 95.11% being the lowest value among all files. Four files were fully classified with 100% specificity. The average sensitivity and specificity values for the Focal class were 91.32% and 89.59%, respectively. Six files were classified with 100% sensitivity and the specificity was consistently higher than 80%, except in one of the files. For the Generalized class, the average sensitivity was 95.23% and the average specificity was 92.63%. Six files of this class were also classified with 100% sensitivity and specificity was above 87% in all files. Grouping the classes with IEDs, the average values for sensitivity and specificity in the classification were 93.28% and 91.11%, respectively.

Occlusion was applied to all the models, yielding examples such as the ones shown in Fig. C.4. One example of each classification outcome is shown for each model, along with the corresponding probability of the epoch containing an IED.

Table 7.1: Average sensitivity (Sens), specificity (Spec), false positive (FP/hour) and true positive rates per hour (TP/hour) for the VGG, ResNet, M1 and M2 models trained with 100:1 weights using Set B (left: training set; right: test set). These values were calculated based on the results of 5-fold cross-validation, using a threshold where the sensitivity is equal to the specificity. The 95% CIs of each parameter are also presented.

	Train				Test			
	Sens	Spec	FP/hour	TP/hour	Sens (%)	Spec (%)	FP/hour	TP/hour
VGG	98.53 (97.61-99.46)	98.69 (97.75-99.63)	22.30 (6.32-38.28)	91.03 (80.35-100.22)	93.05 (88.92-97.19)	93.00 (87.58-98.42)	122.41 (27.63-217.20)	47.72 (45.60-49.84)
Res	94.06 (89.35-95.92)	94.21 (90.02-95.92)	99.03 (17.96-169.71)	86.41 (74.04-95.87)	83.67 (69.12-98.23)	83.20 (70.96-95.44)	293.79 (79.69-507.89)	42.98 (35.43-50.37)
M1	91.06 (83.08-99.05)	91.45 (88.01-94.89)	145.94 (87.30-204.59)	83.55 (70.79-96.31)	82.30 (75.04-89.56)	82.44 (76.53-88.34)	307.11 (203.93-410.39)	42.20 (38.48-45.93)
M2	88.05 (76.94-99.15)	88.57 (77.57-99.58)	195.02 (7.19-382.85)	80.99 (64.47-97.52)	83.10 (72.40-93.79)	82.74 (69.86-95.62)	301.85 (76.55-527.14)	42.61 (37.13-48.10)



(a) VGG (b) ResNet (c) M1 (d) M2

Figure 7.4: Upper row: average ROC curves of the models applied to the training set of Set B; bottom row: average ROC curves of the models applied to the test set of Set B. These were built based on the results of 5-fold cross-validation, with weights 100:1. The 95% CI of the ROC curves is shown as a shaded area. The resulting AUC value and corresponding 95% CIs are also presented.

7.1.3 Normal and Abnormal vs IEDs after removal of normal epochs - Set C

All the models were trained using 100:1 weights for the positive class. Fig. C.5 shows the average ROC curves for the training and test sets after 5-fold cross-validation for the different models trained with weights 100:1, with the 95% confidence interval as a shaded area. The AUC values obtained with the VGG model were 1.00 (CI=1.00-1.00) for the training set and 0.86 (CI=0.83-0.88) for the test set. The ResNet yielded 0.95 (CI=0.92-0.98) on the training set and 0.73 (CI=0.66-0.80) on the test set. The AUC values for M1 and M2 were 0.94 (CI=0.91-0.97) and 0.93 (CI=0.92-0.94) for the training set, with 0.78 (CI=0.75-0.82) and 0.90 (CI=0.89-0.90) having been obtained for the test set.

Table C.2 shows the false positive and true positive rates per hour for the four models at a threshold where the sensitivity equals the specificity. For the VGG, this occurred at 79% sensitivity/specificity for the test set and 85% for the training set. The model led to a false positive rate per hour of 348.60 (CI=111.15-586.06) on the test set and 247.32 (CI=116.88-377.76) on the training set. True positive detections were 49.30 (CI=40.99-57.60) and 41.69 (CI=33.75-49.62) per hour, on the test and training set, respectively. Setting the specificity threshold at 99%, the false positive rates per hour of this model were 23.14 (CI=6.12-40.17) and 17.09 (CI=1.65-32.53), for the training and test set. At 94% sensitivity/specificity, on the training set, the ResNet led to 103.46 (CI=23.00-199.50) false positive detections per hour and 43.56 (CI=39.48-52.28) true positive detections per hour. On the test set, the ResNet achieved 172.4 (CI=29.28-315.53) false positives per hour and 46.18 (CI=41.42-50.94) true positives per hour at a sensitivity of 74% and specificity of 90%. For model M1, false positives per hour were 140.52 (CI=43.29-137.75) and 404.78 (CI=170.19-639.36) on the training and test set, respectively. The true positive rate was 45.11 (CI=38.12-52.10) and 47.82 (CI=39.22-56.42) per hour, at a sensitivity/specificity of 92% for the training set and 76% for the test set. For model M2, this intersection occurred at 88% for the training set and 80% for the test set. False positive detections were 200.43 (CI=57.31-343.57) and 342.32 (CI=119.49-565.16), respectively. The model detected 43.32 (CI=38.50-48.14) true positives per hour on the training set and 49.37 (CI=42.60-56.13) on the test set.

Table C.3 shows the number of epochs, the number of IEDs, the sensitivity and specificity values at a threshold of 0.5, as well as the number of true positives, true negatives, false positives and false negatives at this threshold. The Normal class was classified with an average specificity of 98.47%, with 92.47% being the lowest value among all files. Four files were fully classified with 100% specificity. For the Abnormal class, the average specificity was 97.19%, with one file being classified with 100% specificity and 87.79% being the minimum value on the test set. The average classification specificity in both classes was 95.55%. The average sensitivity and specificity values for the Focal class were 64.09% and 92.67%, respectively. In one of the files, no IEDs were detected, leading to a sensitivity of 0%. For the Generalized class, the average sensitivity was 93.24% and the average specificity was 83.41%. Four files classified with 100% sensitivity and specificity was above 61% in all files. Grouping the classes with IEDs, the average values for sensitivity and specificity in the classification were 78.66% and 88.04%, respectively.

Occlusion was applied to all the models, yielding examples such as the ones shown in Fig. C.6. Two examples of each classification outcome are shown for each model, given the wide diversity of input shapes given to the networks. The corresponding probability of each epoch containing an IED is also presented.

7.2 Focal vs Generalized Epilepsy - Set D

The VGG and the ResNet were trained with equal weights assigned to both classes (assuming Focal IEDs as the positive class). Fig. 7.5 shows the average ROC curves for the training and test sets after 5-fold cross-validation for both models, with the 95% confidence interval as a shaded area. The AUC values obtained with the VGG model were 0.99 (CI=0.99-1.00) for the training set and 0.87 (CI=0.85-0.89) for the test set. The ResNet yielded 0.98 (CI=0.98-0.99) on the training set and 0.79 (CI=0.78-0.80) on the test set.

If a threshold where the sensitivity is equal to the specificity is used, the resulting values for these parameters, as well as for the average of true and false positives and negatives obtained per hour are shown in Table C.4. For the VGG model, this intersection occurred at 96% for the training set and 81% for the test set. False detections were 48.96 (CI=0.00-109.49) and 199.48 (CI=86.94-312.02), respectively. The model detected 491.31 (CI=430.56-552.06) true positives per hour on the training set and 616.62 (CI=509.89-723.36) on the test set. The false negative rate was 18.59 (CI=2.38-34.80) and 147.01 (CI=40.27-253.75) for the training and test set and true negative detections were 1241.1 (CI=1133.20-1349.00) and 813.51 (CI=733.78-893.33). In what concerns the ResNet, on the training set (sensitivity/specificity of 94%), false positive detections were 73.51 (CI=52.95-94.07) per hour and false negative detections were 30.00 (CI=19.84-40.16) per hour; 479.89 (CI=424.86-534.93) true positives and 1216.6 (CI=1146.00-1287.20) true negatives were detected per hour. At 78% sensitivity/specificity, on the test set, these values were 222.86 (CI=143.03-302.69), 161.56 (CI=141.76-181.36), 602.08 (CI=582.28-621.88) and 813.51 (CI=733.78-893.33), respectively.

Occlusion was applied to both models, yielding examples such as the ones shown in Fig. 7.6 . Two examples of each classification outcome are shown for each model, along with the corresponding probability of the epoch belonging to the Focal class.

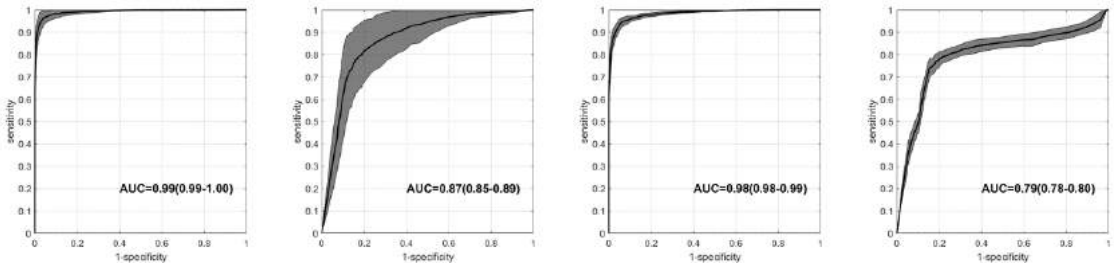


Figure 7.5: Left: ROC curves for the VGG (first panel: training set; second panel: test set). Right: ROC curves for the ResNet (first panel: training set; second panel: test set). These were built based on the results of 5-fold cross-validation. The 95% CI of the ROC curves is shown as a shaded area. The resulting AUC value and corresponding 95% CIs are also presented.

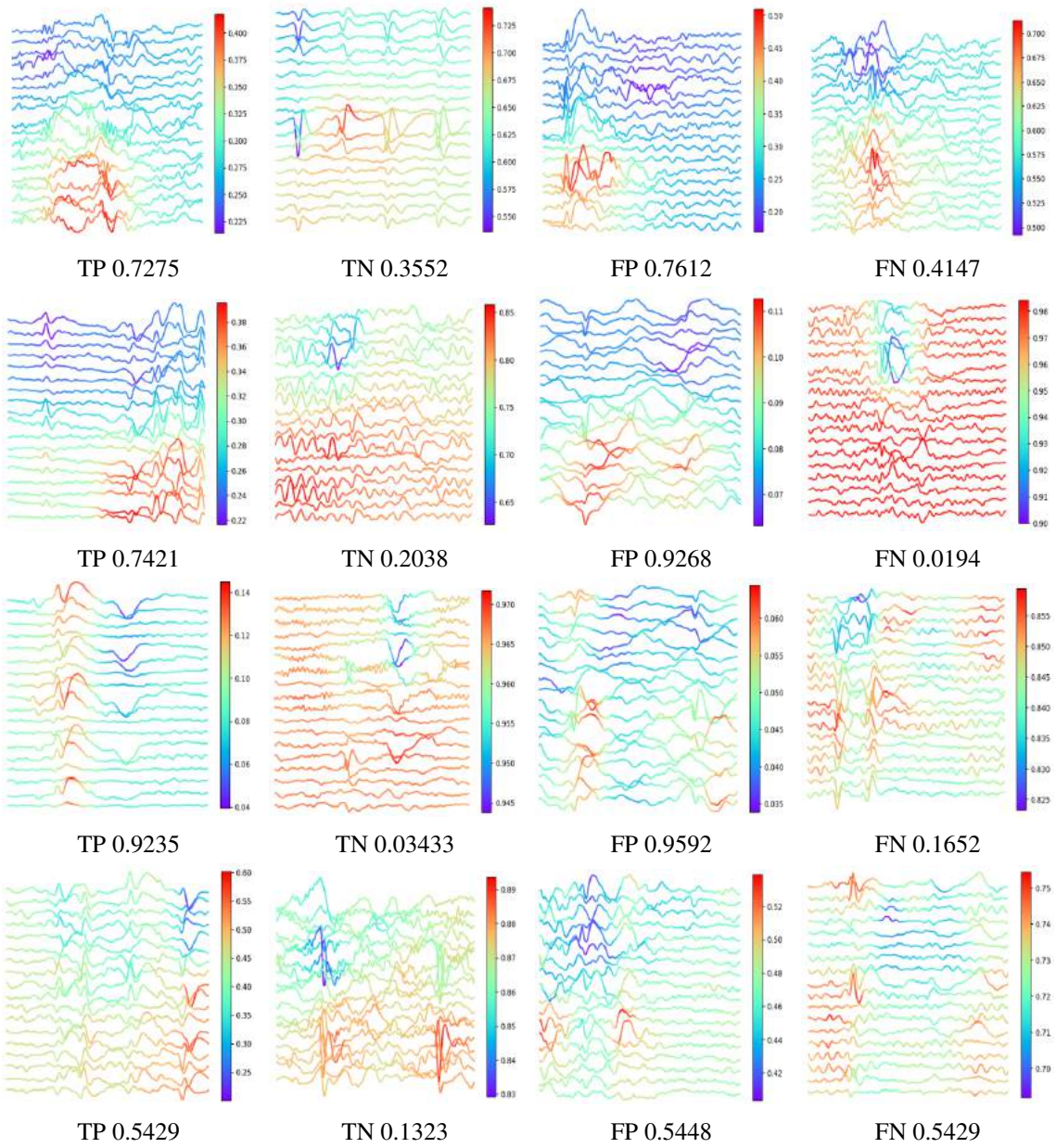


Figure 7.6: Examples of the results obtained with occlusion for the models trained with Set D. First and second rows: VGG; third and fourth rows: ResNet. First column: True Positives; second column: True Negatives; third column: False Positives; fourth column: False Negatives. The probability assigned by the networks regarding the occurrence of an IED in each sample is shown. The scale shows the difference between the probability assigned to the epoch and the probability obtained when a patch is occluded, and warmer colors are assigned to higher differences. Thus, areas plotted in warmer colors are more important for classification.

7.3 Abnormality Detection - Sets E and F

The VGG and the ResNet were trained using Normal data as the negative class and EEGs with non-epileptiform abnormalities as the positive class (Set E), with equal weights assigned to

both classes. The VGG model was also trained using a set which also included IEDs, both Focal and Generalized, in the positive class (Set F).

Fig. C.7 shows the average ROC curves for the training and test sets after 5-fold cross-validation for both models, with the 95% confidence interval as a shaded area. The AUC values obtained without the IEDs in the positive class with the VGG model were 1.00 (CI=1.00-1.00) for the training set and 0.57 (CI=0.46-0.69) for the test set. The ResNet yielded 0.82 (CI=0.67-0.98) on the training set and 0.44 (CI=0.37-0.50) on the test set. When the IEDs were added, the VGG yielded an AUC of 1.00 (CI=1.00-1.00) for the training set and 0.75 (CI=0.72-0.77) on the test set.

Table C.5 shows the false positive and true positive rates per hour for the VGG trained with both sets and for the ResNet trained without IEDs in the positive class, at a threshold where the sensitivity equals the specificity. Trained with Set E, at sensitivity/specificity of 99%, the VGG yielded 9.38 (CI=0.00-21.62) false positive detections per hour on the training set and 322.18 (CI=152.71-491.66) on the test set. The true positive rate was 662.36 (CI=600.92-723.80) and 444.94 (CI=338.61-551.27), respectively, at 64% sensitivity and 70% specificity. Using Set F, the intersection on the training set occurs at the same value, with a false positive rate of 7.53 (CI=1.31-13.65) and true positive rate of 1067.2 (CI=1054.30-1080.10). On the test set, at 74% sensitivity/specificity, there were 169.71 (CI=144.53-194.90) false positive detections per hour and 854.14 (CI=747.03-961.25) true positive detections per hour. For the ResNet model, the false positive rate was 192.37 (CI=117.18-267.56) and the true positive rate was 538.84 (CI=450.75-626.92) on the training set, at a sensitivity of 81% and specificity of 83%. On the test set, at a sensitivity of 48% and specificity of 76%, there were 265.06 (CI=166.28-363.85) false positive detections per hour and 335.86 (CI=279.50-392.23) true positive detections per hour.

7.4 Three-class problem - Sets G and H

The VGG model was trained with these datasets using 3 units in the last fully connected layer. Figs. C.8 and C.9 show the performance of the VGG model when trained using the full EEGs from epilepsy patients and data from the Normal class (Set G) or only IEDs and data from the Normal class (Set H), respectively, using per-class ROC curves.

Using a threshold where the sensitivity is equal to the specificity, Table C.6 shows the accuracy, sensitivity and specificity obtained on the test set of Set G and Set H for each class. The Focal class was detected with an accuracy of 60.05% (CI=51.01-69.09) at a sensitivity/specificity of 60% using Set G and 79.37% (CI=63.73-95.01) at a sensitivity/specificity of 80% using Set H. For the Generalized class, this intersection occurred at 90% on the test set (detection accuracy of 89.99% (CI=82.87-97.11)) with Set H. Trained with Set G, the sensitivity was 57% with a specificity of 55% and accuracy of 56.06% (CI=51.60-60.52). The Normal class was detected with an accuracy of 95.11% (CI=60.79-69.43) (sensitivity/specificity of 65%) using Set G and 82.34% (CI=68.85-95.84) using set H, at a sensitivity of 82% and specificity of 83%.

Fig. 7.7 shows the macroaveraged ROC curves for the training and test set of these datasets, pertaining to all classes. The model achieved an AUC of 0.86 on the training set of Set G and 0.97

on the training set of Set H. On the test set, the VGG yielded AUCs of 0.72 and 0.94 for Set G and H, respectively.

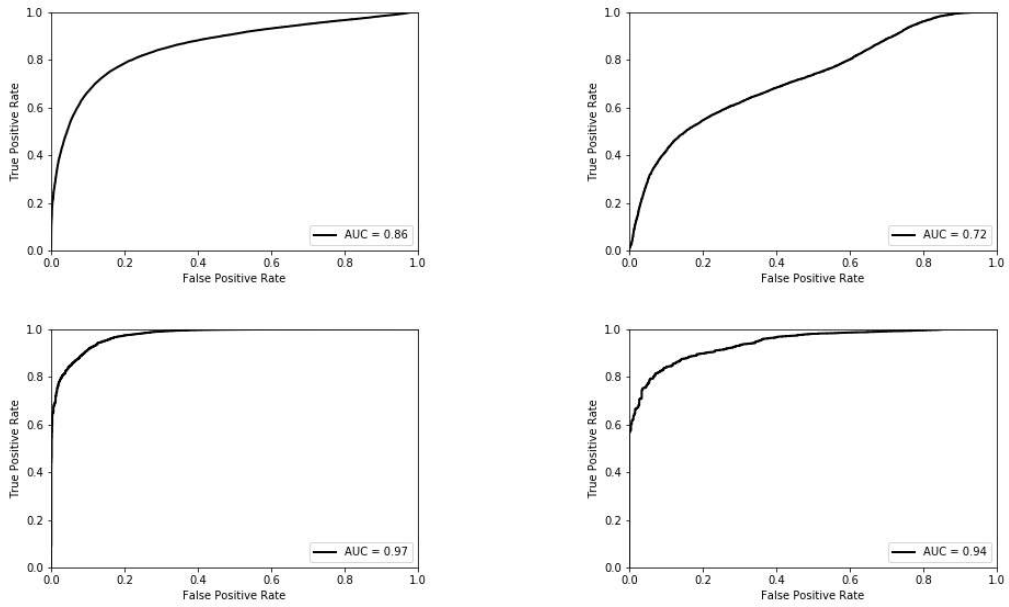


Figure 7.7: Macroaveraged ROC curves for the VGG model trained using Set G (upper row) and Set H (bottom row). The first column shows the results on the training set and the second one presents the results on the test set of these datasets. The AUC value is also showcased.

7.5 Four-class problem - Sets I and J

The VGG model was trained with these datasets using 4 units in the last fully connected layer. Figs. C.10 and C.11 show the performance of the VGG model when trained using sets the full EEGs from epilepsy patients, as well as data from the Normal and Abnormal classes (Set I) or only with IEDs and data from the Normal and Abnormal classes (Set J), respectively, using per-class ROC curves.

Using a threshold where the sensitivity is equal to the specificity, Table C.7 shows the accuracy, sensitivity and specificity obtained on the test set of Set I and Set J for each class. The Focal class was detected with an accuracy of 55.51% (CI=47.42-63.61) at a sensitivity/specificity of 55% using Set I and 77.34% (CI=59.00-95.67) at a sensitivity of 79% and specificity of 77% using Set J. For the Generalized class, this intersection occurred at 67% on the test set (detection accuracy of 67.87% (CI=61.33-74.40)) with Set I. Trained with Set J, the sensitivity was 87% with a specificity of 85% and accuracy of 85.54% (CI=76.01-95.08). The Normal class was detected with an accuracy of 61.11% (CI=57.49-64.74) (sensitivity/specificity of 61%) using Set I and 74.37% (CI=71.26-77.47) using set I, at a sensitivity/specificity of 74%. For the Abnormal class, the intersection was the same using Set J, with an accuracy of 74.76% (CI=72.55-76.96). Using set I, the detection accuracy was 71.17% (CI=58.91-83.43), with sensitivity/specificity of 71%.

Fig. 7.8 shows the macroaveraged ROC curves for the training and test set of these datasets, pertaining to all classes. The model achieved an AUC of 0.89 on the training set of Set I and 0.93 on the training set of Set J. On the test set, the VGG yielded AUCs of 0.65 and 0.82 for Set I and J, respectively.

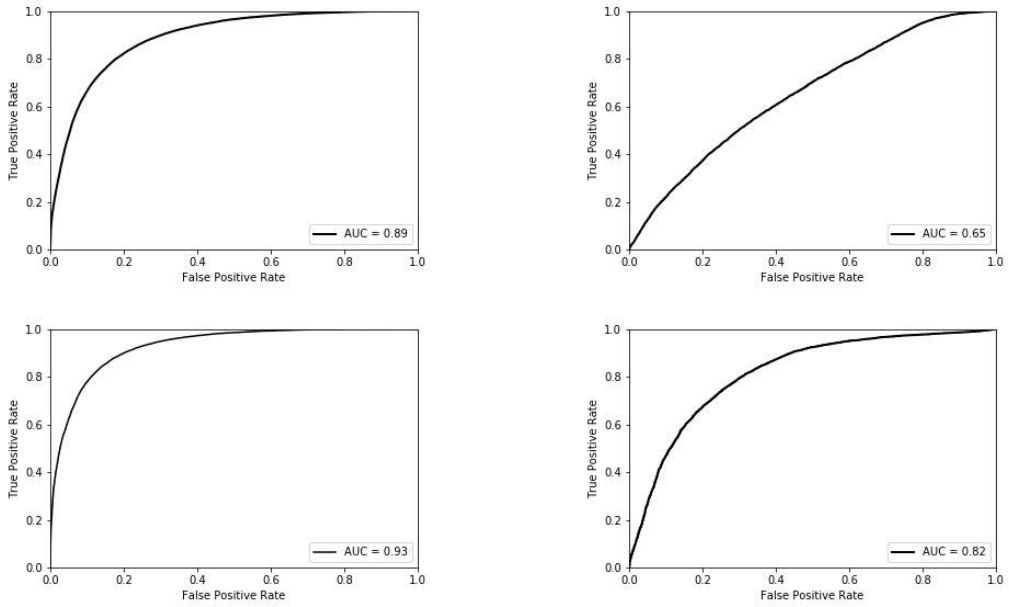


Figure 7.8: Macroaveraged ROC curves for the VGG model trained using Set I (upper row) and Set J (bottom row). The first column shows the results on the training set and the second one presents the results on the test set of these datasets. The AUC value is also showcased.

Chapter 8

Discussion

8.1 IED Detection

8.1.1 Class imbalance and weights

The problem of IED detection was the main issue tackled in this dissertation. The first approach used was training one of the models (in this case, the VGG) with Set A weighing both classes equally. As shown in Fig. C.1, the high accuracy (97.0% on the test set, at a threshold of 0.5) achieved by the model is due to the large number of negative (i.e. Normal) samples when compared to the IEDs (both Focal and Generalized, which amount to approximately 3% of the test set of Set A). In fact, 83% of IEDs were misclassified, as can be seen in the normalized confusion matrix.

To reduce this class imbalance, higher weights were assigned to the positive class (i.e. IEDs). Weights of 10, 50 and 100 were used for this purpose. Fig. C.2 shows that the use of class weights reduces misclassification of the positive class, with true positives rising from 17% without class weights to 90% with weights of 100:1. The number of IEDs being misclassified as Normal (false negatives) decreased from 83% to 10%, at a threshold of 0.5. This showed that the use of class weights was an effective way of reducing class imbalance, allowing the network to learn potentially relevant features for the detection of epileptic discharges.

8.1.2 Performance on Set A

All the models (VGG, ResNet, M1 and M2) were trained using Set A and 100:1 class weights. Since a high AUC value indicates high predictive ability of a model, given that the ROC curve is a tradeoff between sensitivity and the false positive rate (FPR), the AUC values shown in Fig. C.3 prove that all the models were able to distinguish IEDs from normal EEG epochs.

The VGG and M2 yielded the best performances, with both models yielding 0.99 (CI=0.99-0.99) AUC on the training set and 0.91 (CI=0.84-0.94) on the test set of Set A. It is relevant to point out that model M2 and the VGG have a large complexity gap (M2 has 2 only convolutional layers and the VGG includes 13 layers of this type), and, as such, it is interesting to see that they

are able to reach the same performance on this problem, with this dataset. The ResNet and M1 led to lower AUCs: 0.94 (CI=0.85-1.00) and 0.91 (CI=0.88-0.95) on the training set, respectively. On the test set, these models yielded AUCs of 0.76 (CI=0.69-0.84) and 0.70 (CI=0.62-0.78). The confidence intervals were also wider for these models, in particular for the test set, showing a higher uncertainty in the classification process.

Although these results were an improvement from those obtained without class weights, Fig. 7.3 suggests that there were several samples being 'misclassified' by the model (the VGG, in this case) as IEDs (false positives) that were, in fact, IEDs. This happened because, as mentioned in 6.1.1, IEDs in which the experts were in doubt were not annotated, which led to some IEDs being labeled as normal. This means that the actual performance of the networks was potentially better than that shown by the ROCs of Fig. C.3. A way to assess this difference was to remove all the normal epochs from the epileptic EEGs, which led to the creation of Set B.

8.1.3 Performance on Set B

The four networks were retrained using Set B and class weights 100:1 to assess the difference in performance in the IED detection task. Comparing Figs. C.3 and 7.4, it is possible to confirm that, as expected, there was an improvement in the performance of all the models on the test set. The VGG network reached 0.96 (CI=0.95-0.97) on the test set of Set B, followed in performance by M2 (0.91 (CI=0.88-0.94)), ResNet (0.89 (CI=0.84-0.94)) and finally M1 (0.84 (CI=0.83-0.85)).

Table 7.1 shows that the VGG network also achieved the highest intersections of sensitivity and specificity both on the training and test sets (98% and 93%, respectively), with false positive rates of 22.30 (CI=6.32-38.28) and 122.41 (CI=27.63-217.20) per hour on these sets. These were less than half of the next lowest rates, obtained by the ResNet (99.03 (CI=17.96-169.71) and 293.79 (CI=79.69-507.89)), showing that the VGG model was effective in minimizing the number of normal epochs seen as epileptiform discharges. If we set specificity to 99%, as a doctor might want to do in a clinical setting to see only the epochs in which there is a high certainty of the occurrence of an IED, the false positive rate for the VGG becomes 16.50 (4.95-28.06) on the training set and 32.31 (15.15-49.46) on the test set. This shows that an increase in specificity of only 1% on the training set reduces false positives by 5.8 per hour, with roughly one sample being misclassified in each 4 minutes of recording. On the test set, this reduction is almost four-fold when compared to the values on Table 7.1, with one sample being misclassified in each 2 minutes of recording. This model also yielded the highest true positive rates per hour, with 91.03 (CI=80.35-100.22) on the training set and 47.72 (CI=45.60-49.84) on the test set.

This threshold variation can be very useful in clinics, as experts can choose a certain specificity to limit the epochs that have to be manually reviewed, accepting or rejecting the classification of the model. Clinicians can start with the epochs with higher probabilities, where it is more likely that they will indeed find IEDs, allowing them to diagnose patients by looking at a very small number of epochs. If these are not enough for diagnosis, the threshold can be reduced and more epochs will be manually reviewed.

Even with a threshold of 0.5, it is possible to see that every file of the Normal class on the test set is classified with a specificity value over 95%, with several files reaching 100% specificity (i.e. there were no false positives, see Table C.1). This indicates that experts will not be shown many epochs in which the network sees IEDs, if any, if they are analysing a normal EEG, even at a low threshold. In turn, this makes it easier to analyse more EEGs in less time and improve diagnosis.

In what concerns the classification of the classes containing IEDs, it is possible to see in Table C.1 the average sensitivity value is 93% and the average specificity is 91%. The model's performance is slightly better in the Generalized class, which can be explained by the lower number of IEDs in most of the files belonging to the Focal class (one file contained only one IED, for instance). Despite this, the obtained results clearly show that it is possible to use the proposed method to automatically detect IEDs with a high sensitivity and specificity, even at a low threshold.

Notwithstanding the results we obtained, it is not valid to draw a general conclusion regarding the superiority of the VGG model for the IED detection task. While it led to the best results in this experiment, given that all the models were trained using the same set of hyperparameters, it is possible that these were more optimized for this network than for some of the other models, impacting their performance negatively. To assess this, it would be necessary to run hyperparameter optimization for each model separately and compare the results of that experiment.

It is relevant to mention that the use of visualization techniques, in this case occlusion, was able to show that this 'incorrect misclassification' of IEDs was happening in Set A, allowing us to account for human error and improve the models' learning process and, consequently, performance. In fact, it is possible to confirm that this does not continue to happen in set B (cf. Figs 7.3 and C.4). In this case, since all the mislabeled epochs were removed, false positives correspond to samples that were truly misclassified. This is an example of the usefulness and importance of visualization and model interpretability, in particular in a field where human error is abundant.

While these results were satisfactory and proved that the networks were able to detect IEDs, it was still valid to question whether the networks were actually learning to detect IEDs or just abnormalities in general, since these had not been included in these experiments. Set C was then created to answer this question.

8.1.4 Performance on Set C

All the models were trained using Set C and 100:1 class weights to assess if the networks were learning to distinguish IEDs rather than any abnormality. This was done by including non-epileptiform abnormalities in the training set, as part of the negative class.

Model M2 yielded the highest AUC when applied to the test set of Set C (0.90 (CI=0.89-0.90), as seen on Fig. C.5). On the training set, this model led to an AUC of 0.93 (CI=0.92-0.94). The VGG model scored higher on the training set (AUC of 1.00 (CI=1.00-1.00)) but led to an AUC of 0.86 (CI=0.83-0.88) on the test set. This is probably due to the higher complexity of the model, which makes it more prone to overfitting. Model M1 achieved a higher AUC than the ResNet on the test set (0.78 (CI=0.75-0.82) versus 0.73 (0.66-0.80)) but not on the training set

(0.94 (CI=0.91-0.97) versus 0.95 (CI=0.91-0.98)). Despite the low AUC obtained on the test set, Table C.2 shows that the ResNet yielded the lowest false positive rates on both sets, respectively 103.46 (CI=23.00-199.50) and 172.4 (CI=29.28-315.53). If specificity is set at 99%, these values become 0.00 (CI=0.00-0.02) and 17.49 (CI=7.05-27-95). While the False Positive rate is very low on the training set, the value obtained on the test set is slightly higher than what was found for the VGG (17.09 (CI=1.65-32.53)), which suggests some overtraining of the ResNet. While the True Positive rates were rather similar across models, the highest value was obtained by M1 on the training set (45.11 (CI=38.12-52.10)) and by M2 on the test set (49.37 (CI=42.60-56.13)).

Comparing these results with the ones obtained with Set B (cf. Figs. 7.4 and C.5), it is possible to conclude that there was a decrease in the performance of all models. On the test set, the VGG went from an AUC of 0.96 (CI=0.95-0.97) to 0.86 (CI=0.83-0.88) and M2's AUC decreased from 0.91 (CI=0.88-0.94) to 0.90 (CI=0.89-0.90). Comparing Tables 7.1 and C.2, it is possible to see an increase in the false positive rates of all models, further showcasing the decrease in performance. This loss of predictive power is partially due to the increase in difficulty of the problem, since the models now need to learn the differences between epileptiform discharges and other abnormalities. However, it is also caused by the inclusion of more samples (the test set went from 8.8h to 11.4h and the training set increased from 24.3h to 40.5h), which further diluted the (already small) number of IEDs. This means that a possible way to further improve these results is to change (increase) the weights of the positive class to account for this even bigger imbalance. Alternative ways of dealing with this issue, including gathering more data, using synthetic data or data augmentation techniques, will be discussed in Section 9.2.

Taking into consideration the Tables C.1 and C.3, which show the values of sensitivity and specificity at a threshold of 0.5 for each file on the test set for Set B and Set C, respectively, it is possible to see that the average value of specificity in the Normal class remained almost unchanged, with an equal number of files having 100% specificity. The Abnormal class is also classified with high specificity (96%). However, the average sensitivity and specificity of the detection of epileptic classes decreased to 79% and 88%, respectively (versus 93% and 91% on Set B). Again, this was most likely caused by the 'dilution' of IEDs in the new dataset, widening the volume gap between classes.

Despite this decrease in performance, the results yielded by the models on this set were still quite above random. In fact, the results on true positive cases obtained when occlusion was applied to the models trained with this set (see Fig. C.6) clearly showcase the detection of the IED shapes. Thus, it was possible to conclude that the networks were indeed able to distinguish IEDs from other types of abnormalities as well as normal EEGs, meaning that these networks are suitable to aid in the diagnosis of Epilepsy through the detection of epileptiform discharges.

8.1.5 Visualization

Three visualization techniques (filter visualization, input maximization and occlusion) were applied to the VGG network trained with Set A and class weights 100:1.

Filter visualization showed that the VGG's filters differ in what concerns general shape when lower-level layers are compared with higher-level layers. As shown in Fig. 7.1, lower level filters have a somewhat noisy, quite regular pattern across the sample, while filters in higher level layers are active in more specific parts of the signal. These patterns found on filters of higher level layers are closer to an 'IED detector', since their shapes start to resemble epileptiform discharges. In the two panels on the right side of the first row of Fig. 7.1, it is possible to see some activity across channels, suggesting the detection of Generalized discharges. It is also possible to see some isolated patches of activity in the panel on the right, suggesting the detection of Focal discharges. The second row of Fig. 7.1 further illustrates these detection patterns, with the two panels on the left including patches of focal IED detection and the two on the right showing more generalized detection filters. While the first panel on the left has some vertical activity across channels suggesting Generalized detection, the upper channels, on the right side of the filter, show a very clear and isolated detection area that looks like a focal IED. The second panel shows several of these Focal IED-shaped areas, with reduced activity in the remaining filter. The two panels on the left show four and two (respectively) patches of vertical activity (across channels), suggesting the detection of Generalized discharges.

Activation maximization (Fig. 7.2) showed that the parts of the signal that deviate the most from what would be expected in a normal EEG (baseline) led to the highest average activation of all the filters in the VGG. While this makes sense, since these could be parts of the signal associated with abnormalities, the results are not particularly enlightening. On one hand, not all of these areas are highlighted. On the other hand, the highlighted areas have different shapes in different samples (as can be seen by comparing the highlighted areas on the left and right panels of Fig. 7.2), not being directly connected to a 'typical' shape of an IED. However, this is not uncommon, since, as stated by Ancona et al. [330], it is difficult to relate the results obtained with this type of method to a variation of the input. This is particularly true in our case, since the average activation of all the filters is used. In future work, we should change parameters such as the number of iterations and the step size, aiming to make this technique more revealing.

The application of the occlusion technique to the VGG model trained with Set A showed that the network was, in fact, detecting IEDs. This can be concluded since the IED shapes are the patterns highlighted as relevant for classification (see true positive panels of Fig. 7.3). This is a result of the utmost importance, since it proves that the model was learning to perform the desired task in the 'correct' way, i.e. learning features of the epileptiform discharges and not any other spurious characteristics. Furthermore, as mentioned in 8.1.2, this technique also showed that there were mislabeled samples being 'misclassified' by the network as False Positives.

When applied to the four models trained using Set B, occlusion shows that all of them are able to correctly detect IEDs, as these shapes are clearly the relevant areas for the positive classification of a sample (see true positive panels of Fig. C.4). The false positive panels of Fig. C.4 show that the mislabeled epochs were indeed removed, as the misclassified samples are negatives (normal EEG epochs) with parts of the signal or artefacts that are 'seen' as the networks as IEDs. Looking at the false negative panels, it is interesting to see that misclassification happens most commonly

in cases where the epoch from the positive class is noisy or has many IEDs. Finally, it is also possible to see that most detections occur in the second quarter of the epoch, which incites questions regarding whether the networks are more prone to detect IEDs in this area (for instance, due to experts tending to annotate samples approximately 500ms before the discharge occurs). To answer this, a temporal shift should be applied to the samples and these new epochs should be classified by the networks. A correct classification would indicate independence regarding the position of the IED. If incorrect classifications occur (proving the opposite), a way to solve the issue would be to retrain the models with these shifted samples, to make sure the networks have enough examples of IEDs in different parts of the epoch, making their detection independent of this factor.

Set C included samples from EEGs with non-epileptiform abnormalities, introducing a variety of new epoch shapes when compared to the previous sets. These can be seen in Fig. C.6, on some of the true negative and false positive panels, showing that the networks are able to classify some of them correctly (true negatives, mostly chaotic signals) and some of them are getting misclassified (false positives, usually containing some waves losely resembling IEDs). The true positive panels still show that IED shapes are being clearly detecting and used for classification, further proving that the models can learn to identify IEDs instead of abnormalities in a more general sense.

8.1.6 Contextualization in the Literature

As mention in Section 5.4, there have been several attempts to detect IEDs in EEGs of epilepsy patients using a plethora of machine learning methods. However, it is not trivial to compare the performance obtained by different authors due to the diversity of datasets used. Despite this hindrance, it is relevant to assess where our approach stands in the current context. It is, nonetheless, important to keep in mind that all the comparisons drawn in this section are done across datasets, so the differences in performance should not be judged as absolute.

Using template matching, Lodder et al. [281] achieved 90% mean sensitivity and 2.36 false detections per minute. The dataset used in this work included 20-30 minute recordings from 23 epilepsy patients. With data from only 3 patients, Nonclercq et al. [277] were able to get a similar sensitivity value with template matching and k-means. At a sensitivity and specificity of 93%, our false positive rate per minute was 2.04 on the test set, using set B for training, which surpasses the results obtained by these authors. Furthermore, setting the specificity to 99%, our false positive rate per minute becomes 0.5 on the test set, increasing the gap between the performance of our method when compared to [281] and [277].

The combination of ANNs and expert systems has also been used in this problem. Usually, ANNs categorize the inputs and expert systems incorporate spatial context and provide a final classification. Using this type of approach trained with 12h recordings from 7 patients, Argoud et al. [303] achieved 70.78% sensitivity and 69.12% specificity in spike detection, as well as 71.91% sensitivity and 79.19% specificity for sharp waves detection, both of which are below what was obtained using our approach.

Pang et al. [305] compared four different ANN-based methods ([308], [282], [306] and [307]) using the same dataset consisting of 8 channel EEG signals from 7 epilepsy patients and 6 normal

controls. This study revealed that Webber et al.'s algorithm, based on using mimetic features as input for a simple ANN, led to the best performance. It yielded 86.61% sensitivity and 86.32% selectivity, which is still inferior to our performance.

Artameeyanant et al. [296] used a set of 100 single-channel EEGs and developed an approach based on wavelet transform coupled with a neural network trained with parameterized data, which yielded 76.55% sensitivity at 81.30% specificity, coming short of our intersection between sensitivity and specificity at 93% on the test set. Using a similar dataset of 100 single-channel EEGs, Song et al. [300] developed another approach, involving the extraction of the wavelet transform, as well as complexity based features. A genetic algorithm and a neural network were then applied, leading to a sensitivity of 96% and 93.6% specificity. While these values are higher than what was obtained in our experiments, it is necessary to take into account the vast difference between datasets and the fact that the false detection rate is not stated in Song et al.'s work. If this rate is high, the algorithm loses clinical relevance as the experts are still required to review many normal epochs which are being classified as IEDs.

Other neural network-based approaches were developed by Guler et al. [324] and Ubeyli et al. [290,291,312,313]. Guler et al. achieved 96.79% accuracy using Lyapunov exponents as input features for a RNN, with the dataset provided by Andrzejak et al. [323]. Using the same dataset for training, Ubeyli et al. achieved 98.05% accuracy with Lyapunov exponents and a PNN, as well as 99.3% accuracy using eigenvector methods for feature extraction and a SVM classifier. There were more authors experimenting with Andrzejak's dataset. Using a combination of KNN and a genetic algorithm, Guo et al. [245] achieved 93.5% accuracy and Orhan et al.'s [295] approach involving a wavelet transform, k-means and a MLP led to 98.80% accuracy, 99.3% specificity and 98.02% sensitivity. However, apart from Orhan et al's work, the threshold at which the accuracy was calculated is not mentioned. As such, despite these values being higher than the 93% we reached on the test set of Set B, at a threshold of sensitivity/specificity of 93%, it is possible that they were calculated at lower thresholds. Also, the dataset used in these papers is vastly different from the one used in our work, which further complicates this comparison.

Thomas et al. [317] used a CNN as a feature extractor in their approach, followed by an SVM as a classifier. This was trained on 30 minute recordings of 63 controls and 93 epilepsy patients, leading to a 0.935 mean AUC across 4 cross-validation folds, which is lower than our 0.96 (CI=0.95-0.97) obtained on the test set when the VGG is trained with Set B.

Looking at more recent deep learning approaches where neural networks are used as end-to-end classifiers, Johansen et al.'s [316] 5 layer CNN trained on 30 minute EEG recordings from 5 epilepsy patients yielded an AUC of 0.947, just slightly lower than the 0.96 (CI=0.95-0.97) obtained in our experiments. Tjepkema et al. [315] used a set of 50 patients and 50 controls to train a 19 layer, 2D CNN. The results were validated on a set of 5 patients and 12 controls, leading to 0.94 AUC for the test set, with 0.6 False detections per minute at 98% specificity. While these results were close to what we obtained (AUC of 0.96 and 0.5 false detections at 99% specificity), the validation set used in [315] is smaller and thus may lead to larger variability, which means that validation on other/more data could have led to different results.

Apart from the comparison of model performances, which can only be done in a relative way, it is relevant to reinforce that our work is innovative in what concerns the current paradigm due to the application of network architectures from the literature. While architectures such as the VGG and the ResNet have been widely used in other areas, such as image analysis, their use in health-related problems, and in automated EEG analysis in particular, is still scarce. In the narrower scope of IED detection, there are currently no publications using these type of networks. Furthermore, the application of visualization techniques, which provide insight into the networks' behavior and processes, also showcases great innovation. In fact, the application of occlusion allowed us to improve our dataset to account for human errors and it also showed that the networks were detecting IED shapes in the EEG signals, proving the success of our approach. Currently, there are no publications which show the application of these techniques in this type of problem, further showcasing the importance of the developed work.

8.2 Focal vs Generalized Epilepsy

Set D was used to assess if it was possible to train networks to distinguish Focal from Generalized Epilepsy based on IEDs. No class weights were used given that the class imbalance between Epilepsy classes was not very large. The VGG and the ResNet were trained for this purpose, with the VGG leading to the best performance, with an AUC value of 0.98 (CI=0.98-0.99) on the test set and 0.99 (CI=0.99-1.00) on the training set. At 96% sensitivity and specificity, on the training set, it was able to detect 491.31 (CI=430.56-552.06) Focal samples and 1241.1 (CI=1133.20-1349.00) Generalized IEDs per hour. True positive detections were ten times more frequent than false positive detections (in which 'Positive' is equivalent to a Focal IED), with true negatives being 66 times more frequent than false negatives (in which 'Negative' is equivalent to a Generalized IED) on the training set. On the test set, at 81% sensitivity and specificity, the VGG detected 616.62 (CI=509.89-723.36) Focal epochs per hour and 836.88 (CI=724.34-949.42) Generalized IEDs per hour. While the number of samples used in this experiment was quite low, both models performed above random, suggesting that it is possible to distinguish Focal from Generalized IEDs.

Turning to the results of the application of occlusion to both models (see Fig. 7.6), it becomes less clear that this distinction can be performed using the available dataset and the chosen methods. Both Focal and Generalized samples appear to be classified as positive and negative, despite Focal being the positive class. This is due to the fact that the IEDs are labeled per patient and not per discharge. It is not uncommon for patients with Focal epilepsy to have Generalized discharges and vice-versa. If this were to happen in the recordings used in the dataset (which Fig. 7.6 proves it did), the network would be getting IED shapes of both types in both classes. While in a clinical setting this is not critical, since the expert diagnoses the patient based on the patterns found throughout their EEG, taking multiple IEDs and their (dis)similarity into account, in the case of neural networks, this makes it almost impossible to learn robust and discriminant feature of either class.

Given these findings, it becomes impossible to escape the question of how the numerical results contrast with what was shown by occlusion. The most likely answer is that, since the dataset was small, the generalization power of the obtained results is low. While the difference between the results on the training and test set is not too large, an external set could be used to assess if this is the case. A more thought-provoking answer is that the networks may be detecting something other than the IED's shape and using it to classify each epoch as being from a patient with Focal or Generalized epilepsy.

8.3 Normal vs Abnormal EEGs

The first approach used to assess if the models were able to distinguish normal EEG epochs from those with abnormalities was carried out by training the VGG and ResNet with Set E, with no weights assigned to the classes since there was no significant class imbalance. Both models scored high on the training set, with the VGG reaching an AUC value of 1.00 (CI=1.00-1.00) and the ResNet yielding 0.82 (CI=0.67-0.98). However, the results on the test set were quite low, with AUCs of 0.57 (CI=0.46-0.69) for the VGG and 0.44 (CI=0.37-0.50) for the ResNet (see Fig. C.7). This clearly showed overtraining of both models, which were overfitting the training set and not learning robust features. One of the main reasons for this was imposed by the dataset itself: EEGs containing abnormalities are not always abnormal and, since these were not annotated, many normal epochs in these EEGs were being used as part of the positive class, 'confusing' the models.

An attempt to reduce the relative percentage of normal epochs in the positive (i.e. Abnormal) class was made by including Focal and Generalized IEDs in the dataset, creating Set F. The VGG model was trained using this dataset, assessing if it was possible to distinguish normal EEGs from both epileptiform and non-epileptiform abnormalities. The network's performance on the training set remained the same but the AUC value for the test set rose to 0.75 (CI=0.72-0.77), showing improvements regarding what was obtained with Set E. Table C.5 shows that the intersection between the sensitivity and specificity values occurred at a higher threshold on the test set for the VGG trained with Set F (74%), maintaining the same intersection on the training set when comparing to what was obtained with Set E (99%). Thus, overtraining was reduced when training with Set F compared to Set E, although it was still happening.

The total number of IEDs was quite small when compared to the volume of data in the Normal or Abnormal class, so the improvement in performance shown despite this scarcity of 'truly positive' samples shows that it is possible for the network to perform this task successfully, if a correctly labeled dataset is given.

8.4 Multiclass problems

8.4.1 Three-class problem

This three class problem was an extension of the IED detection problem described in 8.1, as it assesses if it is possible to train the VGG to distinguish Focal IEDs from Generalized IEDs and from Normal EEG epochs. We attempted to do this using the full EEG from epilepsy patients (Set G, without removal of the normal epochs from epileptic EEGs) and using only the IEDs of the patients (Set H).

Looking at the macroaveraged ROC curves (Fig. 7.7), it is possible to see that there was a clear improvement when the normal epochs of the epileptic EEGs were removed. Trained using set G, the VGG yielded an AUC of 0.86 on the training set, with this value rising to 0.97 when Set H was used. On the test set, this difference was even larger, with the AUC value going from 0.72 to 0.94. This improvement is also shown in the per-class ROC curves (cf. Figs. C.8 and C.9), with higher AUCs being achieved in all the classes using Set H. Table C.6 is consistent with these results, showing intersections between the sensitivity and specificity values at a higher threshold on Set H, on the test set, for all the classes. The Generalized class showed the highest increase in the threshold, going from a sensitivity of 57% and specificity of 55% to a threshold of 90%, the highest among all the classes. On the other hand, the Focal class had the lowest values of sensitivity (81%) and specificity (79%).

8.4.2 Four-class problem

A further extension of the problem including data from the Abnormal class was explored. The same approaches as described above led to the creation of two analogous datasets (Set I and Set J), with which the VGG was trained.

Consistently with what was observed in the previous section, the macroaveraged ROC curves (Fig. 7.8) and corresponding AUC values improved when the normal epochs were removed from the epileptic EEGs. The AUC score of the model went from 0.89 to 0.93 on the training set and from 0.65 to 0.82 on the test set. While this was still below the 0.93 obtained for the previous problem, a performance decrease was expected due to the introduction of a new class. The problem is even more difficult since it was proven in section 8.3 that distinguishing Normal from Abnormal epochs is quite complex given the lack of annotations in the Abnormal class.

Looking at the per-class ROCs (cf. Figs C.10 and C.11), the improvement from Set I to Set J is also noticeable, with all the AUC values increasing. Table C.7 reinforces this improvement, with intersections occurring at higher values of sensitivity and specificity on the test set of Set J. Once again, the Generalized class had the higher threshold, with 88% sensitivity and 86% specificity on Set J. The Abnormal class led to the lowest intersection value (75%).

Comparing Tables C.6 and C.7, it is possible to see that there was a decrease in the intersection of the sensitivity and specificity values in all the classes. This was more noticeable in the Normal

class (83% on Set H to 74% on Set J), which makes sense given that the new class (i.e. Abnormal) included normal epochs (as shown in Section 8.3), making the learning process more complex.

Still, the obtained results show that the distinction between classes is possible using these methods. However, this could be improved using more IED data (i.e. Focal and Generalized class) and including only abnormal epochs of the Abnormal class (which would require labeling).

8.5 Limitations

The main limitation in these experiments was the scarceness of IED data, which caused class imbalance and made it overall more difficult for the networks to learn relevant features due to the reduced number of samples. The mislabeling of some IEDs as normal epochs can also be seen as a limitation, since it impacted the number of available IEDs and led to 'incorrect misclassifications' by the models.

Another issue that affected some of the experiments (since not all of them included this class) was the lack of labeling of the Abnormal class. This made problems such as those described in sections 8.3 and 8.4.2 much more difficult because normal epochs were being fed to the network as part of both the Normal and the Abnormal classes.

Chapter 9

Conclusions and Future work

9.1 Conclusions

With this work, we show that it is possible to automatically detect Interictal Epileptiform Discharges in EEGs using deep learning methods.

The VGG network yielded the best performance with the set of hyperparameters used, achieving an AUC of 0.96 (CI=0.95-0.97) when distinguishing focal and generalized IEDs from normal EEGs. The false positive detection rate per hour was 122.41 (CI=27.63-217.20) at 93% sensitivity/specificity, dropping to 32.31 (15.15-49.46) at a specificity of 99%. At a 0.5 threshold, the average specificity when classifying normal EEGs was 98.64%, with four EEGs in the test set being classified with 100% specificity. EEG files containing IEDs were classified with an average sensitivity of 93.28% and average specificity of 91.11%. This network was able to reach an AUC of 0.86 (CI=0.83-0.88) when abnormal EEGs were added to the negative class, proving that the model learns to distinguish IEDs and not abnormalities in a general sense. At a threshold of 0.5, the average specificity of detection in the negative class was 95.55%, with EEGs from the positive class being classified with 78.66% average sensitivity and 88.04% average specificity.

Aside from the main question, this dissertation also showed that it is possible to distinguish focal from generalized IEDs using neural networks, even with a very reduced number of samples. We also show that it is possible to distinguish focal IEDs from generalized IEDs and from normal (and abnormal) EEGs.

Furthermore, it was possible to show the relevance of the use of network visualization techniques, as well as the practical applicability of the information they reveal. Filter visualization showed that the filters present in higher level layers of the VGG showcase activity in patches resembling IEDs in isolated channels (focal) and vertically, across channels (generalized). The application of occlusion revealed the presence of mislabeled epochs, allowing performance improvements in later approaches. It also confirmed that IEDs were being clearly identified in true positive cases, proving that the network was learning relevant features.

Despite the limitations in what concerns annotations and the volume of the available data, this dissertation shows the potential of deep learning techniques in this field, in which the automation of

EEG analysis could result in a significant reduction of workload for clinicians, providing objective and reliable results in the diagnosis of Epilepsy. This work is innovative in its use of neural networks described in the literature and in the application of probing techniques to said networks, both of which are not common in healthcare applications and have not been published in the scope of IED detection.

9.2 Future Work

9.2.1 Model Performance

This dissertation was not a self-contained project, but rather the first step towards a larger and more comprehensive take on the problem of IED detection, as well as other deep learning approaches in health (such as prognostication in postanoxic coma). It showed what was already possible given the available data and allowed us to identify limitations and prospective ways to make the models more robust in future experiments.

With that in mind, the first priority in what concerns future work is obtaining more IED data (both focal and generalized). A larger number of samples containing IEDs can reduce class imbalance and provide the networks with more varied and meaningful examples, making the features learned more relevant. The annotation of abnormal EEGs would also be very helpful, particularly if problems in which normal and abnormal EEGs are not in the same class are to be addressed.

Since it is not trivial to obtain very large amounts of IED data, the application of data augmentation techniques and the use of synthetic data are also being considered. Data augmentation can be done by shifting epochs containing IEDs temporally (in the x axis), by adding noise to existing samples or by switching the position of the channels concerning both brain hemispheres. Synthetic data can be generated using mathematical models of IEDs (to which noise can also be added) or using Generative-Adversarial Networks (GANs). It is also possible to train the first layers of the model with synthetic data and then refine this by training higher level layers with patient data, thus requiring less of the latter. Another possible approach is self-supervision, in which data is automatically labeled by finding relations between input patterns.

Trying other simpler models, similar to M1 and M2, can also be a possibility, in particular in problems where overtraining is observed even for a low number of training iterations. Taking blocks from models described in the literature may be an option to decide on the architecture of these simpler models. Understanding the lower results of the ResNet when compared to the VGG in the IED detection task would also be of interest, since this can be due to the chosen hyperparameters but also to the types of filters, skip connections and other characteristics of the architecture.

Other types of networks such as recurrent neural networks (LSTMs and others) which take into account the temporal relationship between samples could also be an alternative approach to this problem. Another possibility is to apply networks from the literature with pre-trained weights. While the datasets used in this problem are far from those which were originally used to train the

networks, it could be interesting to see how much this initial training can impact the problem at hand and potentially improve performance.

Since clinicians often resort to different montages when analysing EEGs, it would be interesting to train networks with data from several montages and see if one of them led to better results, or even if adding a final classifier, creating an ensemble model, could improve the overall performance. Another way to experiment with ensemble models would be to build binary classifiers for each class and join them to perform the final classification. Yet another idea would be, if enough data was available, to train per patient classifiers, since IEDs from a single patient are usually similar in shape. These could then be assessed on their own or combined, allowing comparisons to be drawn with the patient-independent classifier. Furthermore, attempting to cluster IEDs per shape instead of patient (thus including several patients per cluster) and building classifiers based on said clusters, could also improve performance.

While it was possible to prove that networks use spatial information in their classification process in what concerns IEDs, it would be interesting to see what the models could discern based on single-channel EEG. While this approach could be more appropriate for other problems (such as postanoxic coma prognostication, in which the EEG pattern in analysis is usually consistent across channels), it would make the process of synthesizing data from mathematical models easier. It would also be possible to combine the results from per-channel classifiers, incorporating some spatial notions (albeit not as strongly as when the model is trained with the full EEG).

Another interesting question concerns the spatial gap in the bipolar montage, used in these experiments, and how it can affect model performance. In this montage, there is an abrupt spatial transition from the right to the left hemisphere of the brain, breaking the spatial continuity of the electrodes. Since 2D networks take spatial context into account, it would be of interest to suppress this discontinuity. A possible way to do this is by using 3D networks and 'plotting' each electrode's position in one of the dimensions of the input.

Some other ideas that can be explored in future work are the optimization of the deep learning models' hyperparameters, as well as the train/test split ratio. The use of an external validation set, ideally with data from a different medical center, would greatly add to the proof of robustness of the trained models.

9.2.2 Visualization

The visualization techniques applied in this dissertation were able to reveal crucial information that enforced confidence in the models' decisions and suggested ways to improve the training process. However, improving these techniques and building on them is paramount to extend the current understanding of the networks, the way they learn and, consequently, classify samples.

Both occlusion and input maximization have parameters such as the number of iterations, step size and window size that can be optimized. In what concerns window size in occlusion, experimenting with single channel windows as well as vertical windows could provide information regarding how the networks see focal and generalized IEDs, respectively. Also, several approaches described in the literature use grey masks for occlusion instead of black (or an average of the

image/signal instead of zero), which could also impact the results obtained with these techniques. Furthermore, obtaining the results of occlusion for different networks working on the same task, using the same signal as input, can showcase differences between the models and the way they look at the EEGs and the IEDs in particular.

Creating a prototypical pattern of an IED based on the network's perception of the discharge itself would show how the model sees and identifies the IED. This prototype could also be used for training, serving as synthetic data. Furthermore, if clustering of IED types is successful, getting prototypes of each cluster using the models would be even more interesting, as it would showcase the differences in the perception of the network of these different shapes.

Exploring other visualization approaches described in the literature (see section 4.5) and combining them can aid in creating a robust framework that allows us to probe deeper into the networks and gain further insights.

Appendix A

State of the Art of Machine Learning in IED Detection

Table A.1: State of the Art of Machine Learning in IED Detection

Ref	Author	Year	Dataset	Method	Results
[44]	Gotman	1976	16 channel recordings from 30 normal subjects and 63 epilepsy patients (2min each)	Mimetic method based on features such as relative amplitude and duration of both half-waves, relative sharpness, total duration	605 sparks/sharp waves (SSW) detected in 30mins of recording of epileptic patients, 1 artefact detected as SSW; 4 SSWs detected in 30mins of recordings of normal subjects, 3 artefacts detected as SSW
[231]	Gotman	1982	16 channel recordings from 24 epilepsy patients (6h recordings)	See [44]	Average of 41 valid and 39 false SSW detections in 6h
[266]	Oliveira	1983	4 channel EEG from 10 epilepsy patients	Mimetic method based on features such as sharpness of the peak, steepness and duration	When compared to 8 electroencephalographers (EEGers), the specificity of the classifier was higher than the lowest of the EEGers, but the sensitivity was lower

[264]	Gotman	1991	20 100min recordings from epilepsy patients, 16 channel EEG	State analysis + mimetic method to classify the recording in one of 5 states (active wakefulness, quiet wakefulness, desychronized EEG, phasic EEG and slow EEG) and apply an adaptation of the detection algorithm described in [44]	16.2% error in the state classification
[265]	Gotman	1992	See [264]	See [264]	16.5% error in the state classification; 43% and 41% true detections on the training and test set, respectively; 52% and 56% false detections on the training and test set, respectively
[331]	Gabor	1992	8 channel EEG from 5 epilepsy patients	ANN (3 layers) with the spatial distribution of the average voltage slopes used as input	Recognition of unequivocal epileptiform complexes with a sensitivity of 94.2% ($\pm$7.3%)
[267]	Hostetler	1992	6 20min recordings from epilepsy patients, 16 channel EEG	Mimetic method based on the ratio of spike and averaged background amplitudes	89% consistency in its detections, surpassing 83% of the EEGers to which it was compared
[272]	Sankar	1992	11 recordings from epilepsy patients, 16 channel EEG	Parametric method (autocorrelation) + template matching	The maximum percentage of real epileptic patterns detected was 17.5%
[268]	Dingle	1993	11 recordings from epilepsy patients, 16 channel EEG	Mimetic method to detect candidate patterns in a single channel + expert system to detect multichannel events	45-71% epileptiform events were detected at 100% selectivity, 60-100% were detected with up to 9 false detections per hour
[269]	Pietila	1994	12 recordings from 6 epilepsy patients	Mimetic method using features such as amplitude and the second derivative of the signal + template matching	97% sensitivity at 78% specificity in files with clear spike-slow-wave bursts, 31% sensitivity at 33% specificity in the others
[282]	Webber	1994	2 to 3min recordings from 10 epilepsy patients, 49 channel EEG	Mimetic method + ANN (3 layers, fully connected) with either parameterized or raw EEG data used as input	Intersection between sensitivity and selectivity occurred at 73% for parameterized data and at 46% for raw data
[306]	Kalayci	1995	16 channel EEG from 5 epilepsy patients	Wavelet transform (Daubechies 4 and Daubechies 20) + ANN (3 layers, 3 to 8 neurons as input for the hidden layer)	The best result was 90.3% accuracy, 87.3% sensitivity at 93.3% specificity when using Daubechies 20 and 8 neurons for the hidden layer
[270]	Benlamri	1997	137 recordings of normal subjects and epilepsy patients, 8 and 16 channel EEG	Mimetic method + expert system to classify epileptiform events based on parameters such as orientation, synchrony, amplitude and duration	87% accuracy (119 recordings were correctly classified)

[301]	Feucht	1997	19 channel EEG from 10 epilepsy patients	Hilbert transform + ANN (MLP was compared to other classifiers such as LDA and cascade-correlation neural networks)	84.6% mean selectivity, 88.1% mean sensitivity, 89.3% mean specificity; all classifiers showed the same performance in terms of false detections, with MLP reaching the highest number of correctly identified spikes
[286]	Park	1998	16 channel EEG from 32 epilepsy patients	Wavelet transform (Daubechies 4) + ANN (3 layers, with the output of the wavelet transform as input) + expert system	97% sensitivity ar 89.5% selectivity for parameterized data
[307]	Ozdamar	1998	16 channel EEG from 5 epilepsy patients	ANN (trained with backpropagation, with varying number of input layers and nodes)	Best performance was achieved with 30 input nodes and 6 hidden layers, with 86% and 82% true classification rates for the training and test sets, respectively
[308]	Tarassenko	1998	20 channel recordings from 2 epilepsy patients (31 and 77 recordings, respectively)	Time and frequency analysis (mobility, complexity, autoregressive modeling) + ANN (three layers, with features as input)	85.6-95.6% accuracy, 83.1-97.3% sensitivity, 85.9-95.5% specificity in spike detection for patient-specific classifiers
[283]	James	1999	16 channel EEG from 15 epilepsy patients for training, 7 recordings from epilepsy patients and one normal EEG for testing	Mimetic method, using features such as amplitude, duration, slope and sharpness + ANN (self-organizing feature map) + fuzzy logic to combine spatial information	55.3% sensitivity at 82% selectivity, 7 false detections per hour on the test set
[284]	Wilson	1999	20 channel EEG recordings from 50 epilepsy patients (40 for training, 10 for validation) and 10 control subjects	Mimetic method + ANN (monotonic neural network)	89.9% sensitivity, 80.1% selectivity, 99.6% specificity on the validation set
[332]	Goelz	2000	11 recordings, total 278min, 16 channel EEG	Wavelet transform (continuous wavelet transform with the complex-valued psi-1 wavelet) + thresholding	84% sensitivity, 12% selectivity
[333]	Calvagno	2000	36 recordings of 8 channel EEG	Wavelet transform + smoothed nonlinear energy operator (SNEO)	Showed the possibility of detecting spikes with the proposed method
[271]	Black	2000	521 recordings, average 20min, 16 channel EEG	Mimetic method + expert system to integrate spatial and temporal information	76% sensitivity, 41% selectivity, 0.41 false detections per hour

[327]	Ko	2000	300 single channel EEG recordings with an IED and 300 without (2/3 for training and 1/3 for testing) from 20 patients	ANN (three layers, 30 nodes in the input layer and 6 in the hidden layer, using raw data as input)	Performance below random
[309]	Kurth	2000	32 channel EEG from 4 patients	ANN (Kohonen feature map)	80.2% average sensitivity and selectivity at crossover threshold
[334]	Nuh	2002	8 channel EEG data divided into 2.56s segments	ANN (wavelet neural network)	82.6% detectability, 90.4% selectivity
[287]	Liu	2002	8 channel EEG from 81 epilepsy patients, over 800h of recordings	Wavelet transform + ANN (MLP with input features such as amplitude, duration, sharpness and slope) + expert system to reject artefacts	98% correct detection of sharp transients, false detection rate of 6.1%
[285]	Castellaro	2002	2000 EEG traces	Mimetic method + ANN (MLP with three or four layers, with 32 output nodes) + expert system	94% of properly classified traces by the ANN, 80% by the expert system
[328]	Latka	2003	19 channel EEG from 1 epilepsy patient	Wavelet transform (mexican hat) + thresholding	70% sensitivity, 67% selectivity
[305]	Pang	2003	8 channel EEG from 7 epilepsy patients and 6 normal controls	ANN (compared methods developed by Tarassenko [308], Webber [282], Kalayci [306] and Ozdamar [307])	Webber's algorithm led to the best performance, with 86.61% sensitivity and 86.32% selectivity
[335]	Durka	2004	EEGs from an online database, with data split into 4 groups (single spikes and sharp waves, test signals, series of spikes, artefacts)	Parametric representation (matching pursuit) + thresholding	92% sensitivity, 84% selectivity
[336]	Adjouadi	2004	20-30min recordings from 18 epilepsy patients	Walsh transform	85% precision, 79% sensitivity, 7.2 false detections per hour
[318]	Acir	2004	19 channel EEG from 25 epilepsy patients (18 for training, 7 for testing)	Parametric representation for pre-classification into spike candidates and trivial non-spikes + SVM	90.3% sensitivity, 88.1% selectivity, 9.5% false detection rate
[310]	Nigam	2004	200 single channel EEG segments (100 from normal controls, 100 from epilepsy patients), 23.6s	Non-linear pre-processing to extract relative amplitude and frequency + ANN (LAMSTAR neural network, uses self-organising maps, features used as input)	1.6% miss rate, 97.2% accuracy

[319]	Acir	2005	19 channel recordings from 29 epilepsy patients (19 for training, 10 for testing)	ANN (2 perceptrons with 6 features as input) to classify EEG peaks into definite IEDs, definite non-IEDs and possible IEDs + SVM to separate the third group	89.1% sensitivity, 85.9% selectivity, 7.5 false detections per hour
[314]	Srinivasan	2005	200 single channel EEG segments (100 from normal controls, 100 from epilepsy patients), 23.6s	Extraction of morphological features such as dominant frequency, average power and normalized spectral entropy + ANN (Elman RNN)	99.6% accuracy with a single input feature
[288]	Guler	2005	Andrzejak's dataset	Wavelet transform (Daubechies 2) + adaptive neuro-fuzzy inference system (ANFIS, 20 extracted features from the wavelet transform as input)	98.68% accuracy
[324]	Guler	2005	Sets A, D and E from Andrzejak's dataset	Lyapunov exponents + ANN (Elman RNN compared with MLP)	Best performance was achieved by the RNN, leading to 96.79% accuracy
[304]	Tzallas	2006	5s EEG recording from an epilepsy patient	Parametric appoach (time-varying autoregressive model with parameters estimated via Kalman filtering) + thresholding	The parametric approach reached better results than SNEO
[337]	Xu	2006	EEG recordings from a normal subject and 8 epilepsy patients	Morphological filtering (average weighted combination of open-closing and close-opening operation) + thresholding	91.62% detection rate
[303]	Argoud	2006	12h EEG recordings from 7 epilepsy patients	Wavelet transform (coiflet wavelet function) + ANN (4 networks to detect 4 types of patterns: spikes, sharp waves, ocular artefacts and noise) + expert system	70.78% sensitivity, 69.12% specificity in spike detection; 71.91% sensitivity, 79.19% specificity for sharp waves
[338]	Exarchos	2006	16 channel recordings from 12 normal controls and 13 epilepsy patients, 15min each	Mimetic method for transient detection + feature selection + association rule mining	87.38% accuracy, 77.5-91.3% sensitivity, 91.75-99.23% specificity, 83.10-93.55% selectivity
[339]	Tzallas	2006	16 channel recordings of 12 normal controls and 13 epilepsy patients	ANN to classify segments into spikes, muscle activity, eye blinks or sharp alpha activity (trained with 16 features as input) + expert system	84.44% accuracy
[340]	Xu	2007	EEG recordings from 2 normal subjects and 10 epilepsy patients	Morphological filtering to separate background activity from spikes (average weighted combination of open-closing and close-opening operation) + thresholding	This type of morphological filtering led to less false detections (7.52%) than traditional morphological filtering (20.48%) and mexican-hat wavelet functions (16.72%)

[289]	Guler	2007	Andrzejak's dataset	Wavelet transform + Lyapunov exponents + SVM compared to ANNs (probabilistic neural network and MLP, using features as input)	75.6%, 72.0%, 68.8% accuracy for SVM, PNN and MLP, respectively
[302]	Adeli	2007	Set of healthy controls, interictal and ictal periods from Andrzejak's dataset	Wavelet transform + Lyapunov exponents and correlation dimension	Correlation dimension is discriminant for higher frequency (beta and gamma subbands), LLE is discriminant for the alpha band
[341]	Inan	2007	19 channel EEG from 8 epilepsy patients	ANN for pre-classification + fuzzy C-means clustering compared to k-means	Fuzzy C-means with pre-classification led to the best results, 93.3% sensitivity, 74.1% specificity
[273]	El-Gohary	2008	32 channel EEG from 2 epilepsy patients	Template matching	96% sensitivity, 4.8 false detections per hour
[342]	Indiradevi	2008	EEG recordings from 22 epilepsy patients	Wavelet transform (Daubechies 4) + thresholding	91.7% sensitivity, 89.3% specificity, 78.1% selectivity, 90.5% accuracy
[290]	Ubeyli	2008	Andrzejak's dataset	Eigenvector methods for feature extraction + SVM compared to ANN (MLP) with features used as input	SVM achieved the highest accuracy (99.3%)
[312]	Ubeyli	2008	Andrzejak's dataset	Eigenvector methods for feature extraction + ANN (probabilistic neural network (PNN) compared to MLP) trained with features as input	PNN achieved the highest accuracy (97.63%)
[311]	Ubeyli	2008	Andrzejak's dataset	Wavelet transform + ANN (Mixture of experts (ME) compared with MLP)	ME achieved the highest accuracy (93.17%)
[343]	De Lucia	2008	21 channel EEG from 7 epilepsy patients	ICA and PCA + mixture of Gaussians	ICA led to the best performance, yielding 65±22% sensitivity at 86±7% specificity
[344]	Keshri	2009	Single channel EEG, 4min recordings	Deterministic Finite Autodata	95.68±3.22% accuracy
[345]	Kutlu	2009	19 channel EEGs	Deterministic Finite Autodata	95.68±3.22% accuracy
[346]	Guo	2009	100 EEG segments of normal subjects and 100 of epilepsy patients	Wavelet transform + ANN (3 layers, 16 input neurons, 10 in the hidden layer, 1 in the output)	98.17% sensitivity, 92.12% specificity, 95.2% accuracy
[291]	Ubeyli	2009	Andrzejak's dataset	Eigenvector methods for feature extraction (power spectral density) + RNN compared with MLP	RNN achieved the highest accuracy (98.15%)

[325]	Ubeyli	2009	Sets A, D and E from Andrzejak's dataset	Wavelet transform + ANN (combined MLP)	94.83% accuracy
[274]	Vijayalakshmi	2010	Single channel EEG segments	Template matching + thresholding	Template matching can provide good results for spike detection
[313]	Ubeyli	2010	Andrzejak's dataset	Lyapunov exponents + ANN (probabilistic neural network (PNN), compared with MLP, trained with features as input)	PNN achieved the highest accuracy (98.05%)
[320]	Lima	2010	Sets A and E from Andrzejak's dataset	Wavelet transform (Daubechies 4) + SVM and least-squares SVM (LS-SVM)	Both types of SVM achieved good discrimative power when trained with raw data, as well as with features
[321]	Kelleher	2010	16 channel EEG from 8 generalized epilepsy patients, 20-40mins	SVM with gaussian kernel, trained with features as input with 5-fold cross-validation	93.47% ROC for the patient-specific classifier
[292]	Abibullaev	2010	Andrzejak's dataset	Wavelet transform (Daubechies 2,4; Biorthogonal 1.3,1.5) + ANN (MLP)	95.49% sensitivity, 93.80% specificity, 94.69% accuracy
[347]	Boos	2011	35 EEG segments of sharp waves, spikes, background activity, alpha waves, artefacts and blinks	Extraction of morphological features + ANN (MLP with 3 layers, 7 to 11 neurons in the hidden layer, with features as input)	90% correct identification
[275]	Ji	2011	21 channel EEGs from 2 epilepsy patients, 46-50min	Template matching	Patients may have different patterns in different channels, so multi-channel templates may be adequate
[348]	Juozapavicius	2011	Several 1h EEG recordings	Morphological filtering (open-close-close-open) + thresholding	Recognized epileptic spikes but also spikes not related to epilepsy, as well as false positives in noisy areas
[276]	Ji	2011	19 channel EEG recordings	Template matching	Automatic detection is a useful assistant tool
[293]	Haydari	2011	Data from [328]	Genetic algorithm + wavelet transform (Daubechies 4) + thresholding	96% sensitivity, 88.8% selectivity for an optimal threshold of 33%
[245]	Guo	2011	Andrzejak's dataset	Wavelet transform + genetic algorithm + KNN	93.5±1.2% accuracy for the genetic programming + KNN algorithm, 67.2±1.2% for the KNN alone
[294]	Wang	2011	Andrzejak's dataset	Wavelet transform + KNN + expert system	100% accuracy with 5-fold cross-validation

[326]	Iscan	2011	Andrzejak's dataset	Cross correlation method to extract time features, power spectral density to extract frequency features + SVM, LV-SVM KNN, Parzen window classifier, LDA, Decision tree, Naive Bayes, Nearest mean classifier, Quadratic classifier	LV-SVM, binary decision trees and quadratic classifiers led to highest accuracies; combination of time and frequency features led to higher accuracy
[322]	Martinez-Vargas	2011	Andrzejak's dataset	Time-frequency analysis (parametric and non-parametric) + KNN	All time-frequency approaches showed similar outcomes, with accuracies ranging from 96% to 99%
[295]	Orhan	2011	Andrzejak's dataset	Wavelet transform (Daubechies 2) + K-means + ANN (MLP, using the output of k-means as input)	98.80% accuracy, 99.33% specificity and 98.02% sensitivity in the diagnosis of epilepsy
[277]	Nonclercq	2012	EEG recordings from 3 epilepsy patients, 20-30min	Template matching + K-means	90.6% sensitivity, 89.9% selectivity
[296]	Artameeyanant	2012	6 groups of 100 single channel EEG segments of 23.6s (spike, epileptic, eyes closed, eyes opened, body movement, normal)	Wavelet transform + ANN (trained using features such as approximate entropy, derived from the wavelet transform, as input)	76.55% sensitivity, 81.30% specificity, 89.47% accuracy
[297]	Sezer	2012	EEG recordings from 240 normal controls and 240 epilepsy patients	Wavelet transform (Daubechies 2) + ANN (MLP, Elman network, general regression network, probabilistic network)	General regression network getting 100% sensitivity, specificity, selectivity and accuracy on a random test set
[298]	Suresh	2013	EEG recordings from normal controls and epilepsy patients	Wavelet transform (Daubechies 4) + energy estimation + ANN (trained using the features as input)	Spikes were successfully detected
[281]	Lodder	2013	20-30min recordings from 23 epilepsy patients; 8 used for training, 15 for testing	Template matching	Mean sensitivity of 90% with 2.36 false positives per minute
[279]	Liu	2013	16 channel EEGs from 3 normal controls and 12 epilepsy patients	Nonlinear energy operator (k-NEO) + Mimetic method + Adaboost classifier	87.4-93.9% accuracy, 87.9-95.5% sensitivity, 86.7-92.4% specificity for spike detection in the test set
[299]	Halford	2013	100 30s EEG recordings from 100 epilepsy patients	Wavelet transform + Fourier transform + bayesian classifier + ANN	Average sensitivity of 58.4% and specificity of 68.3%, over all the classifiers and feature sets
[300]	Song	2013	100 single channel EEG recordings from 5 normal controls and 5 epilepsy patients, 23.6s each	Wavelet transform (Daubechies 4) + extraction of complexity based features + genetic algorithm + ANN (extreme learning machine (ELM))	96.0% sensitivity, 93.6% specificity, 94.8% accuracy

[349]	Radmehr	2013	EEG from 1 epilepsy patient	Wavelet transform + parametric method (time-varying autorregressive model (TVAR)) + thresholding	It was possible to detect spikes
[350]	Chavakula	2013	75 EEG samples during sleep, 8 during wakefulness	Wavelet transform + thresholding + SVM compared to Bayesian classifier	SVM led to the best ROC curve; it was possible to detect spikes in all stages of the sleep cycle and wakefulness
[351]	Zhou	2013	21 channel EEGs from 100 patients (30s each)	Wavelet transform (Daubechies 2 and 4) + KNN (k=3, 10-fold cross-validation)	Wavelet features improve classification in up to 5.75% in sensitivity and 6.76% in specificity
[13]	Lodder	2014	8 EEGs from epilepsy patients used for training, 15 for testing	SVM + template matching	95% of IEDs were detected after 15 iterations
[352]	Janca	2014	EEG recordings from 30 patients with refractory epilepsy	Envelope modeling + thresholding	97.4±12.2% sensitivity, 36.3±22.7% selectivity, 8.2±7.4 false positive detections per minute
[353]	Horak	2015	12-24 channel EEG from 9 epilepsy patients	Template matching + SVM	Template matching led to higher agreement with experts when compared to template matching + SVM
[354]	Chaibi	2015	2min segments of stereotactic EEG	Wavelet transform (continuous and discrete) + thresholding	Discrete wavelet transform leads to better performance in low signal to noise ratios (85.50% sensitivity)
[316]	Johansen	2016	30min recordings from 5 epilepsy patients	ANN (1D CNN with 5 layers)	AUC of 0.947
[355]	Thomas	2016	30min recordings from 50 epilepsy patients	K-means + K-medoids + fuzzy C-means clustering + agglomerative clustering + affinity propagation template matching	The affinity propagation-based template matching system led to the highest AUC (0.953)
[280]	Jing	2016	30min recordings from 100 epilepsy patients	Dynamic time warping (DTW) + Template matching	The algorithm reduces the time spent on annotating spikes in about 70%
[315]	Tjepkema	2018	50/12 EEGs from normal controls and 50/5 from focal epilepsy patients for training/testing	ANN (1 and 2D CNNs and LSTMs)	0.94 AUC for the test set; 47.4% sensitivity, 98.0% specificity, 0.6 false detections per minute
[356]	Bagheri	2018	EEGs from 63 normal controls and 93 epilepsy patients (average 28.5min)	SVM (single SVM compared to SVM cascade, 5-fold cross-validation)	The cascade method increases precision, decreasing false positive detections
[317]	Thomas	2018	30min EEG recordings from 63 normal controls and 93 epilepsy patients	ANN (CNN, used for waveform classification) + SVM (4-fold cross-validation)	83.86% accuracy, 55% precision at 80% sensitivity, 0.935 mean AUC

Appendix B

Supplementary Figures of Chapter 6 - Methods

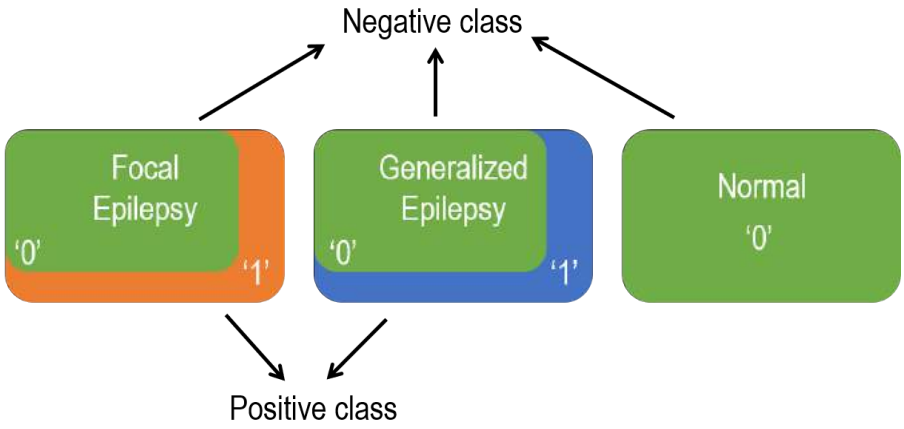


Figure B.1: Schematic view of Set A.

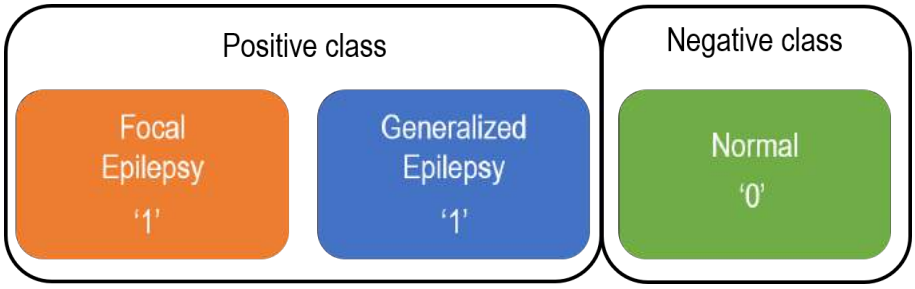


Figure B.2: Schematic view of Set B.

Figure B.3: Schematic view of Set C.

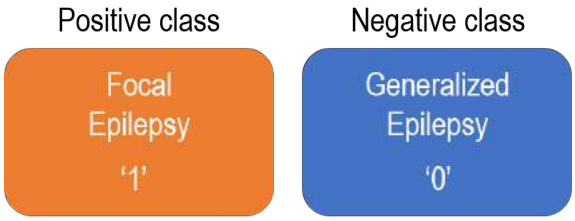


Figure B.4: Schematic view of Set D.

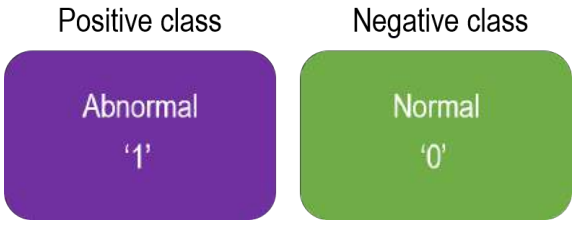


Figure B.5: Schematic view of Set E.

Figure B.6: Schematic view of Set F.

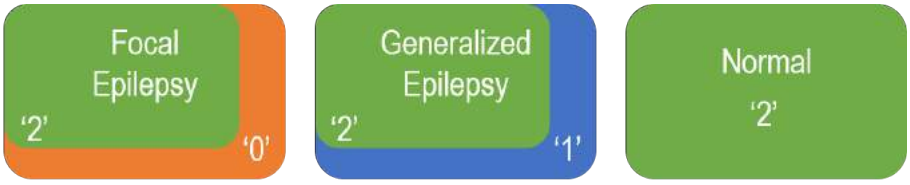


Figure B.7: Schematic view of Set G.

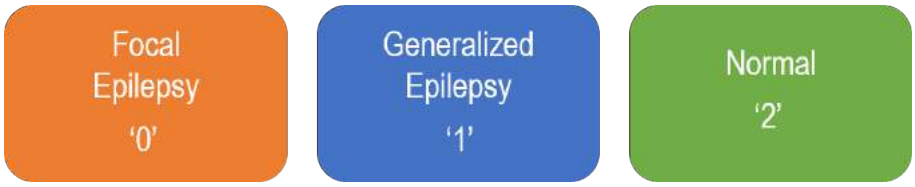


Figure B.8: Schematic view of Set H.

Figure B.9: Schematic view of Set I.

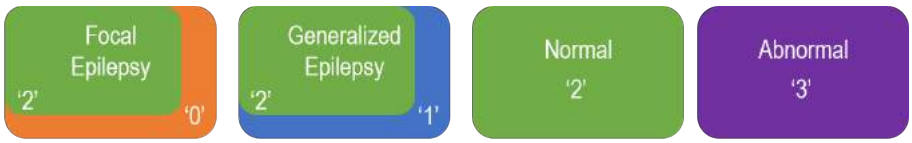


Figure B.10: Schematic view of Set J.

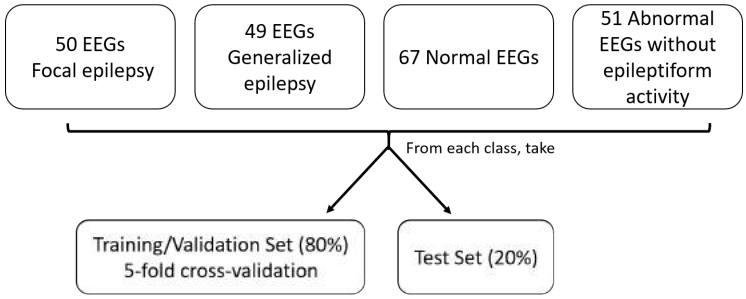


Figure B.11: Separation of the data into the train/validation and test sets.

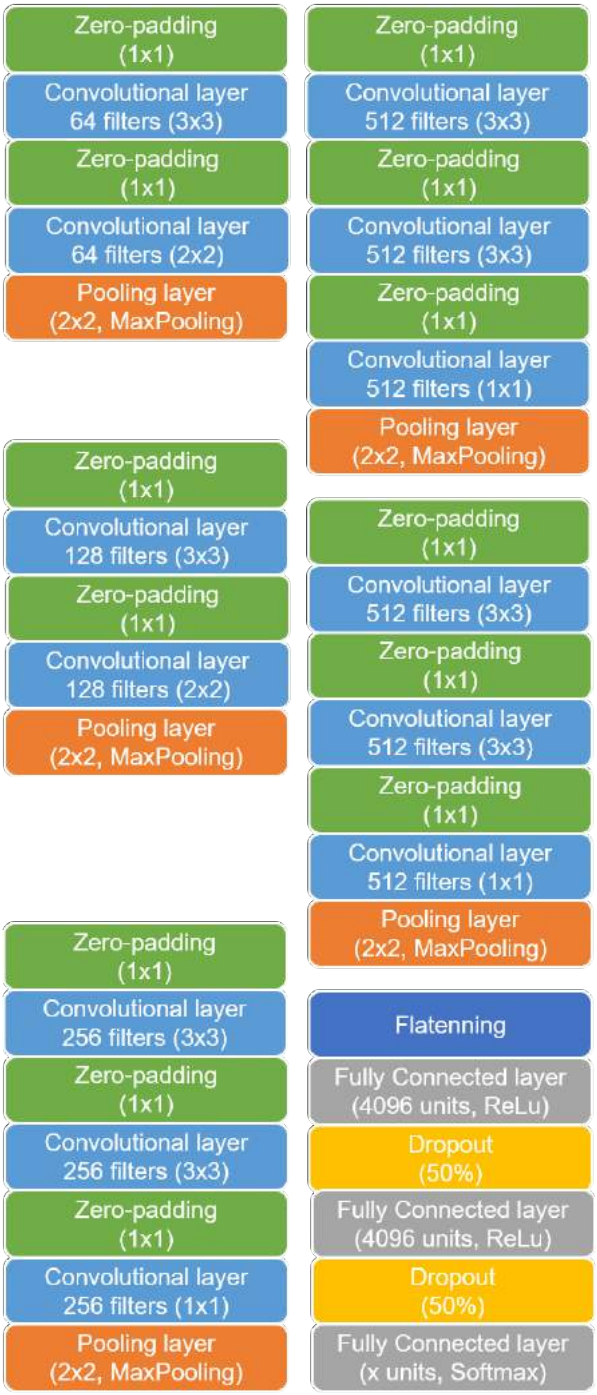


Figure B.12: Architecture of the altered VGG C model.

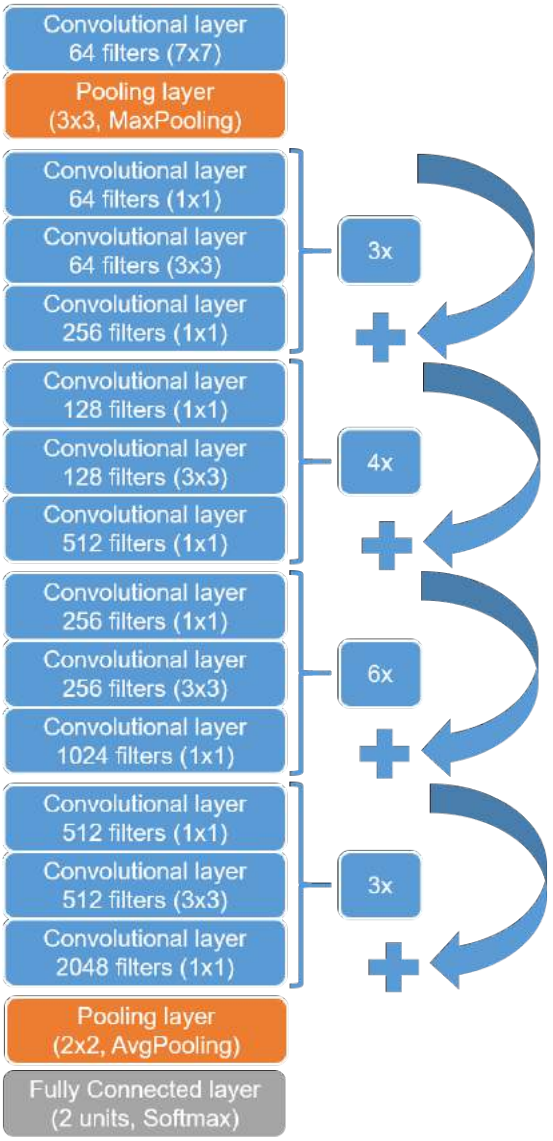


Figure B.13: Architecture of the altered ResNet50 model.

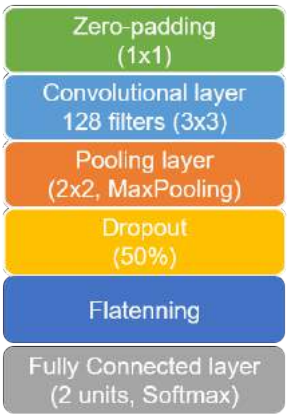


Figure B.14: Architecture of the M1 model.

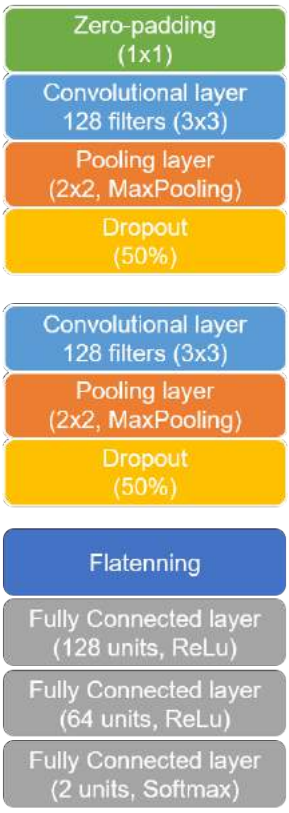


Figure B.15: Architecture of the M2 model.

Appendix C

Supplementary Figures of Chapter 7 - Results

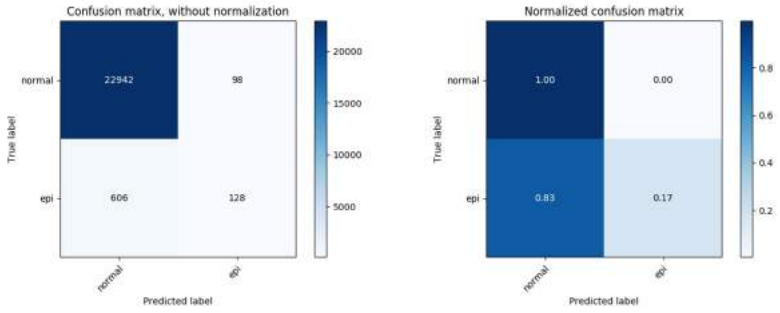


Figure C.1: Left: confusion matrix for the VGG network applied to the test set of Set A, with threshold set at 0.5, without weights. Right: normalization of the confusion matrix shown on the left.

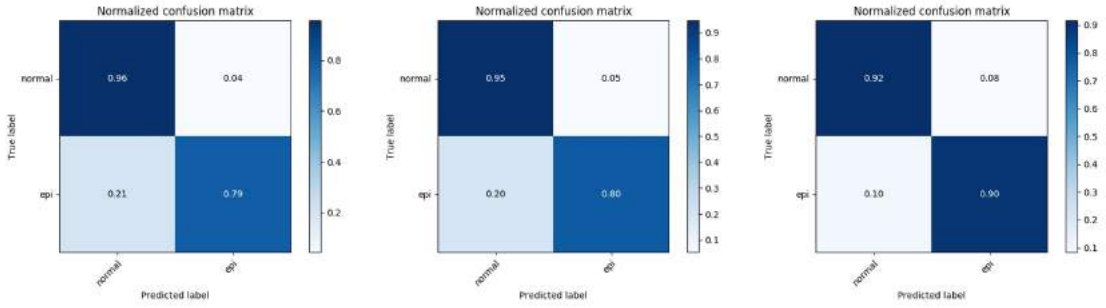
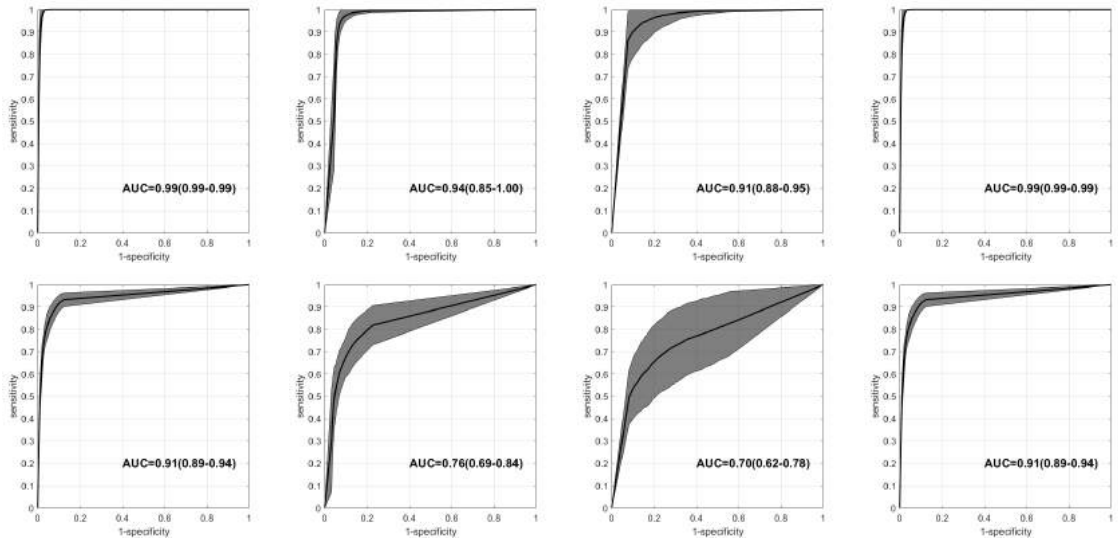


Figure C.2: Normalized confusion matrix for the VGG network applied to the test set of Set A, with threshold set at 0.5, with weights 10:1 (left), 50:1 (center) and 100:1 (right).



(a) VGG (b) ResNet (c) M1 (d) M2

Figure C.3: Upper row: average ROC curves of the models applied to the training set of Set A; bottom row: average ROC curves of the models applied to the test set of Set A. These were built based on the results of 5-fold cross-validation, with weights 100:1. The 95% CI of the ROC curves is shown as a shaded area. The resulting AUC value and corresponding 95% CIs are also presented.

Table C.1: Number of epochs (Epochs), number of IEDs (IEDs), Sensitivity (Sens), Specificity (Spec), True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) in each recording on the test set of set B, classified by the VGG with weights 100:1, at a threshold of 0.5.

	Epochs	IEDs	FP	FN	TP	TN	Sens (%)	Spec (%)
Normal	603	0	0	0	0	603	-	100.00
	538	0	6	0	0	532	-	98.88
	673	0	0	0	0	673	-	100.00
	748	0	34	0	0	714	-	95.45
	643	0	1	0	0	642	-	99.84
	598	0	15	0	0	583	-	97.49
	753	0	2	0	0	751	-	99.73
	2053	0	16	0	0	2037	-	99.22
	2067	0	51	0	0	2016	-	97.53
	2760	0	135	0	0	2625	-	95.11
	1521	0	5	0	0	1516	-	99.67
	598	0	12	0	0	586	-	97.99
	1273	0	0	0	0	1273	-	100.00
	586	0	0	0	0	586	-	100.00
Focal	280	117	86	5	112	77	95.73	47.24
	1154	39	32	12	27	506	69.23	94.05
	1162	3	16	0	3	562	100.00	97.23
	204	3	9	1	2	192	66.67	95.52
	583	15	10	0	15	558	100.00	98.24
	2903	65	134	12	53	2704	81.54	95.28
	665	3	4	0	3	658	100.00	99.40
	614	1	14	0	1	559	100.00	97.72
	663	10	125	0	10	528	100.00	80.86
	613	10	58	0	10	545	100.00	90.38
Generalized	618	43	70	1	42	505	97.67	87.83
	594	8	29	0	8	557	100.00	95.05
	592	9	33	0	9	550	100.00	94.34
	590	9	18	2	7	563	77.78	96.90
	626	4	24	0	4	598	100.00	96.14
	631	34	25	3	31	572	91.18	95.81
	195	4	40	0	4	151	100.00	79.06
	604	21	32	3	18	551	85.71	94.51
	588	36	62	0	36	490	100.00	88.77
	638	3	13	0	3	622	100.00	97.95

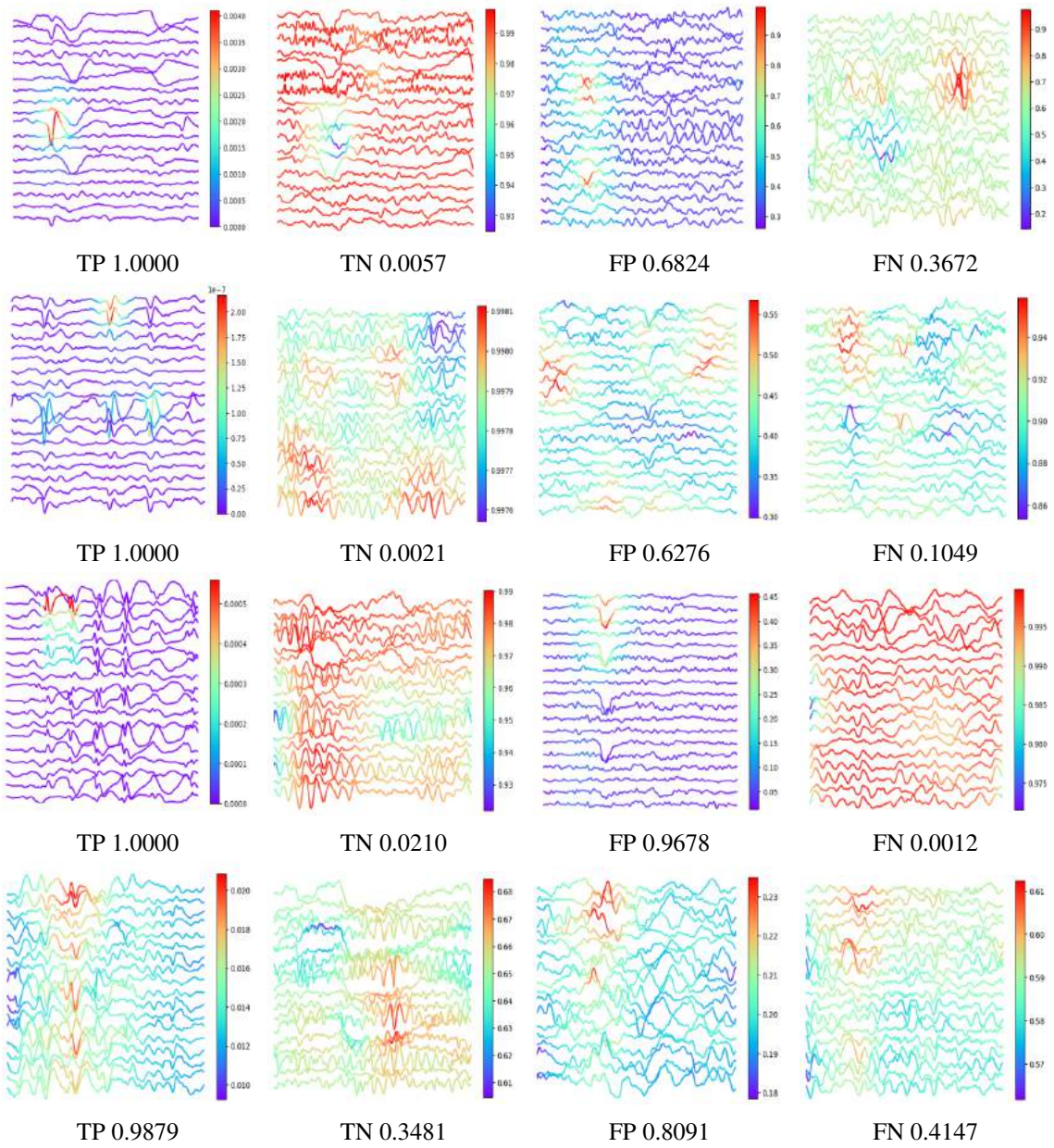
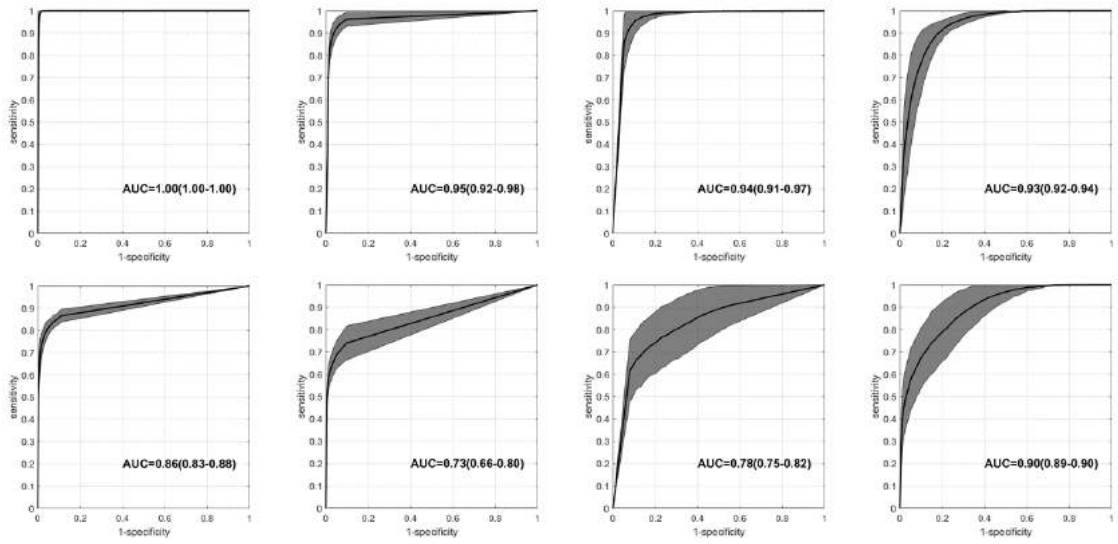


Figure C.4: Examples of the results obtained with occlusion for the models trained with Set B and weights 100:1. First row: VGG; second row: ResNet; third row: M1; fourth row: M2. First column: True Positives; second column: True Negatives; third column: False Positives; fourth column: False Negatives. The probability assigned by the networks regarding the occurrence of an IED in each sample is shown. The scale shows the difference between the probability assigned to the epoch and the probability obtained when a patch is occluded, and warmer colors are assigned to higher differences. Thus, areas plotted in warmer colors are more important for classification.

Table C.2: Average sensitivity (Sens), specificity (Spec), false positive (FP/hour) and true positive rates per hour (TP/hour) for the VGG, ResNet, M1 and M2 models trained with 100:1 weights using Set C (left: training set; right: test set). These values were calculated based on the results of 5-fold cross-validation, using a threshold where the sensitivity is equal to the specificity. The 95% CIs of each parameter are also presented.

	Train				Test			
	Sens (%)	Spec (%)	FP/hour	TP/hour	Sens (%)	Spec (%)	FP/hour	TP/hour
VGG	85.04 (76.44-93.65)	85.87 (78.40-93.34)	247.32 (116.88-377.76)	41.69 (33.75-49.62)	79.07 (65.75-92.39)	79.94 (66.27-93.60)	348.60 (111.15-586.06)	49.30 (40.99-57.60)
Res	93.84 (89.92-97.76)	94.05 (88.61-98.68)	103.46 (23.00-199.50)	43.56 (39.48-52.28)	74.08 (66.44-81.71)	90.08 (81.84-98.32)	172.4 (29.28-315.53)	46.18 (41.42-50.94)
M1	92.32 (83.74-1.00)	91.97 (86.40-97.54)	140.52 (43.29-137.75)	45.11 (38.12-52.10)	76.70 (62.91-90.49)	76.71 (63.21-90.21)	404.78 (170.19-639.36)	47.82 (39.22-56.42)
M2	88.91 (80.84-96.99)	88.56 (80.42-96.70)	200.43 (57.31-343.57)	43.32 (38.50-48.14)	79.18 (68.33-90.04)	80.30 (67.48-93.12)	342.32 (119.49-565.16)	49.37 (42.60-56.13)



(a) VGG (b) ResNet (c) M1 (d) M2

Figure C.5: Upper row: average ROC curves of the models applied to the training set of Set C; bottom row: average ROC curves of the models applied to the test set of Set A. These were built based on the results of 5-fold cross-validation, with weights 100:1. The 95% CI of the ROC curves is shown as a shaded area. The resulting AUC value and corresponding 95% CIs are also presented.

Table C.3: Number of epochs (Epochs), number of IEDs (IEDs), Sensitivity (Sens), Specificity (Spec), True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) in each recording on the test set of set C, classified by the VGG with weights 100:1, at a threshold of 0.5.

	Epochs	IEDs	FP	FN	TP	TN	Sens (%)	Spec (%)
Normal	793	0	4	0	0	789	-	99.50
	533	0	9	0	0	524	-	98.31
	563	0	5	0	0	558	-	99.11
	673	0	0	0	0	673	-	100.00
	603	0	7	0	0	596	-	98.84
	663	0	5	0	0	658	-	99.25
	603	0	0	0	0	603	-	100.00
	593	0	0	0	0	593	-	100.00
	2053	0	69	0	0	1984	-	96.64
	1895	0	26	0	0	1869	-	98.63
	1723	0	23	0	0	1700	-	97.34
	598	0	45	0	0	553	-	92.47
	1273	0	0	0	0	1273	-	100.00
	610	0	9	0	0	601	-	98.52
Abnormal	623	0	53	0	0	570	-	91.49
	658	0	0	0	0	658	-	100.00
	533	0	12	0	0	521	-	97.75
	393	0	48	0	0	345	-	87.79
	668	0	20	0	0	648	-	97.01
	429	0	24	0	0	399	-	94.33
	477	0	4	0	0	473	-	99.16
	637	0	4	0	0	633	-	99.37
	512	0	20	0	0	492	-	96.09
	633	0	19	0	0	614	-	97.00
	981	0	88	0	0	893	-	91.03
Focal	577	39	35	12	27	503	69.23	93.49
	523	95	22	60	35	406	36.84	94.86
	611	5	11	1	4	595	80.00	98.18
	883	11	47	4	7	825	63.64	94.61
	534	67	65	6	61	402	91.04	86.08
	606	3	12	0	3	591	100.00	98.01
	639	53	1	53	0	585	0.00	99.83
	550	43	42	23	20	465	46.51	91.72
	581	23	164	3	20	394	86.96	70.61
	612	6	4	2	4	602	66.67	99.34
Generalized	618	43	85	0	43	490	100.00	85.22
	594	8	27	1	7	559	87.50	95.39
	654	20	182	0	20	452	100.00	71.29
	489	122	101	14	108	266	88.52	72.48
	195	4	74	0	4	117	100.00	61.16
	589	11	36	0	11	542	100.00	93.77
	604	21	33	5	16	550	76.19	94.34
	1121	34	116	2	32	971	94.12	89.33
	614	12	14	1	11	588	91.67	97.67
	510	89	112	5	84	309	94.38	73.40

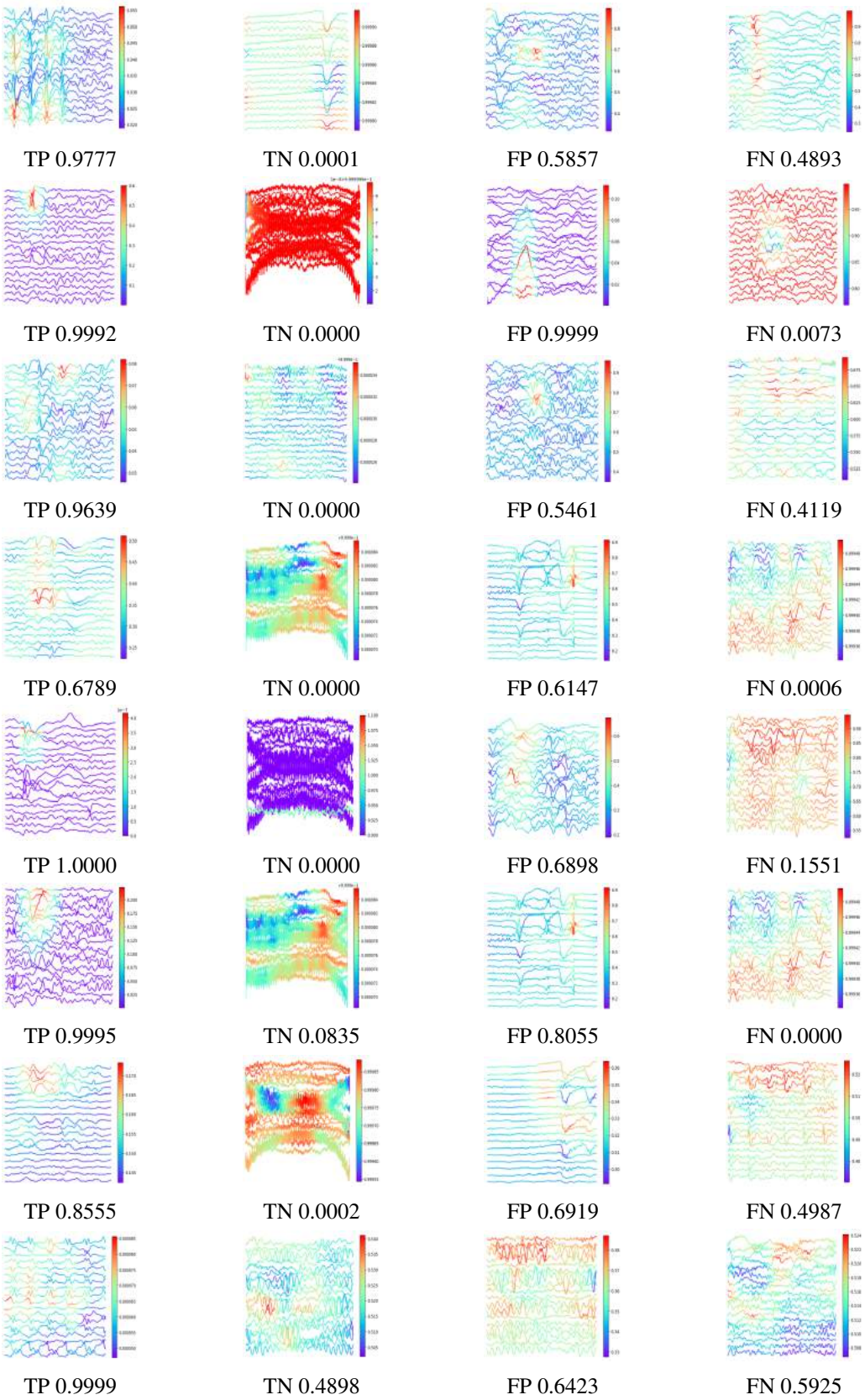
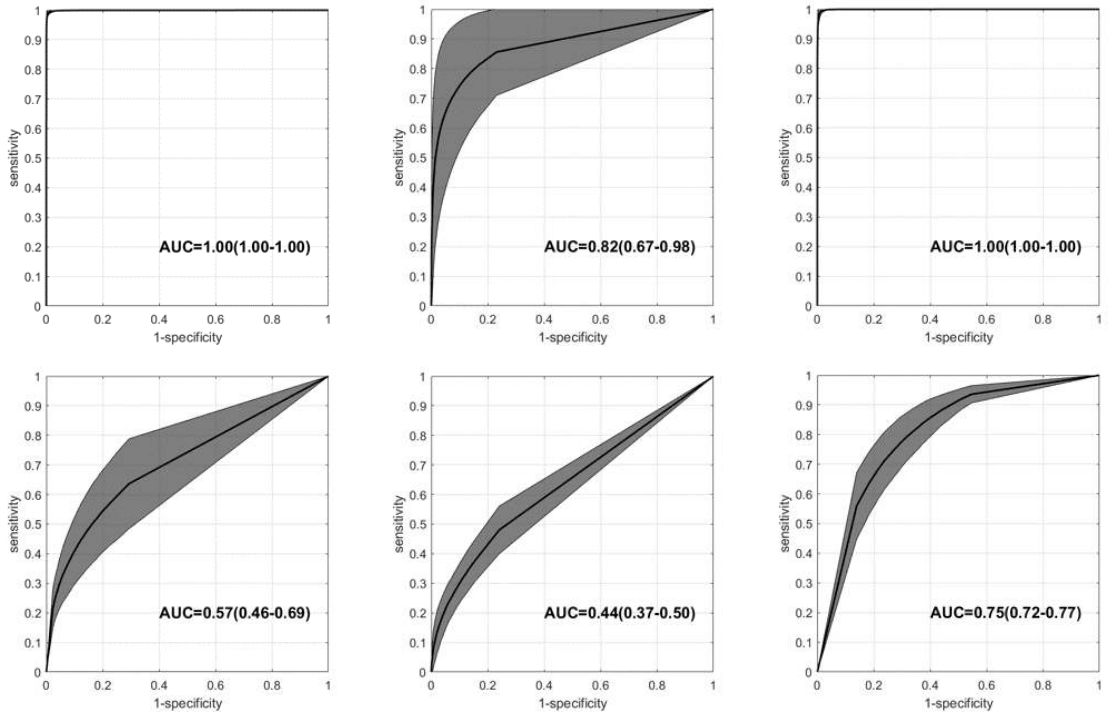


Figure C.6: Examples of the results obtained with occlusion for the models trained with Set B and weights 100:1. First and second rows: VGG; third and fourth rows: ResNet; fifth and sixth rows: M1; seventh and eighth rows: M2. First column: True Positives; second column: True Negatives; third column: False Positives; fourth column: False Negatives. The probability assigned by the networks regarding the occurrence of an IED in each sample is shown. The scale shows the difference between the probability assigned to the epoch and the probability obtained when a patch is occluded, and warmer colors are assigned to higher differences. Thus, areas plotted in warmer colors are more important for classification.

Table C.4: Average sensitivity (Sens), specificity (Spec), false positive (FP/hour), false negative (FN/hour), true positive (TP/hour) and true negative rates (TN/hour) per hour for the VGG and ResNet models trained with Set D (top: training set, bottom: test set). These values were calculated based on the results of 5-fold cross-validation, using a threshold where the sensitivity is equal to the specificity. The 95% CIs of each parameter are also presented.

	Train					
	Sens (%)	Spec (%)	FP/hour	FN/hour	TP/hour	TN/hour
VGG	96.26 (92.91-99.60)	96.12 (91.24-1.00)	48.96 (0.00-109.49)	18.59 (2.38-34.80)	491.31 (430.56-552.06)	1241.1 (1133.20-1349.00)
Res	94.05 (91.8-96.21)	94.27 (92.45-96.09)	73.51 (52.95-94.07)	30.00 (19.84-40.16)	479.89 (424.86-534.93)	1216.6 (1146.00-1287.20)
	Test					
	Sens	Spec	FP/hour	FN/hour	TP/hour	TN/hour
VGG	80.75 (66.77-94.73)	80.75 (69.89-91.61)	199.48 (86.94-312.02)	147.01 (40.27-253.75)	616.62 (509.89-723.36)	836.88 (724.34-949.42)
Res	78.84 (76.25-81.44)	78.5 (70.79-86.20)	222.86 (143.03-302.69)	161.56 (141.76-181.36)	602.08 (582.28-621.88)	813.51 (733.78-893.33)

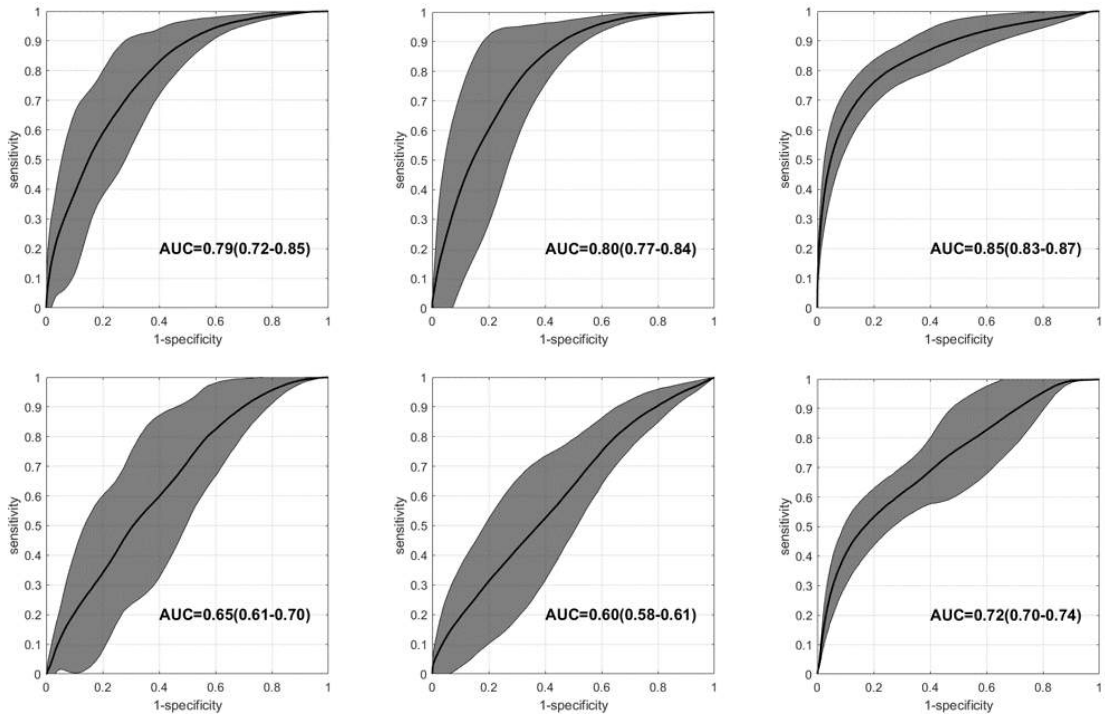


(a) VGG, set E (b) ResNet, set E (c) VGG, set F

Figure C.7: Upper row: average ROC curves of the models applied to the training set of Sets E/F; bottom row: average ROC curves of the models applied to the test set of Sets E/F. These were built based on the results of 5-fold cross-validation. The 95% CI of the ROC curves is shown as a shaded area. The resulting AUC value and corresponding 95% CIs are also presented.

Table C.5: Average sensitivity (Sens), specificity (Spec), false positive (FP/hour) and true positive rates per hour (TP/hour) for the VGG model trained with Set E and Set F and for the ResNet model trained with Set E(left: training set; right: test set). These values were calculated based on the results of 5-fold cross-validation, using a threshold where the sensitivity is equal to the specificity. The 95% CIs for each parameter are also presented.

	Train				Test			
	Sens	Spec	FP/hour	TP/hour	Sens (%)	Spec (%)	FP/hour	TP/hour
VGG (Set E)	99.18 (98.43-99.93)	99.19 (98.16-100.00)	9.38 (0.00-21.62)	662.36 (600.92-723.80)	63.62 (48.42-78.83)	70.73 (55.33-86.13)	322.18 (152.71-491.66)	444.94 (338.61-551.27)
ResNet (Set E)	81.33 (63.79-98.86)	83.04 (76.67-89.41)	192.37 (117.18-267.56)	538.84 (450.75-626.92)	48.03 (39.97-56.08)	75.92 (66.94-84.89)	265.06 (166.28-363.85)	335.86 (279.50-392.23)
VGG (Set F)	98.96 (97.63-100.00)	98.96 (98.11-99.81)	7.53 (1.31-13.65)	1067.2 (1054.30-1080.10)	73.92 (64.65-83.18)	73.66 (69.76-77.57)	169.71 (144.53-194.90)	854.14 (747.03-961.25)

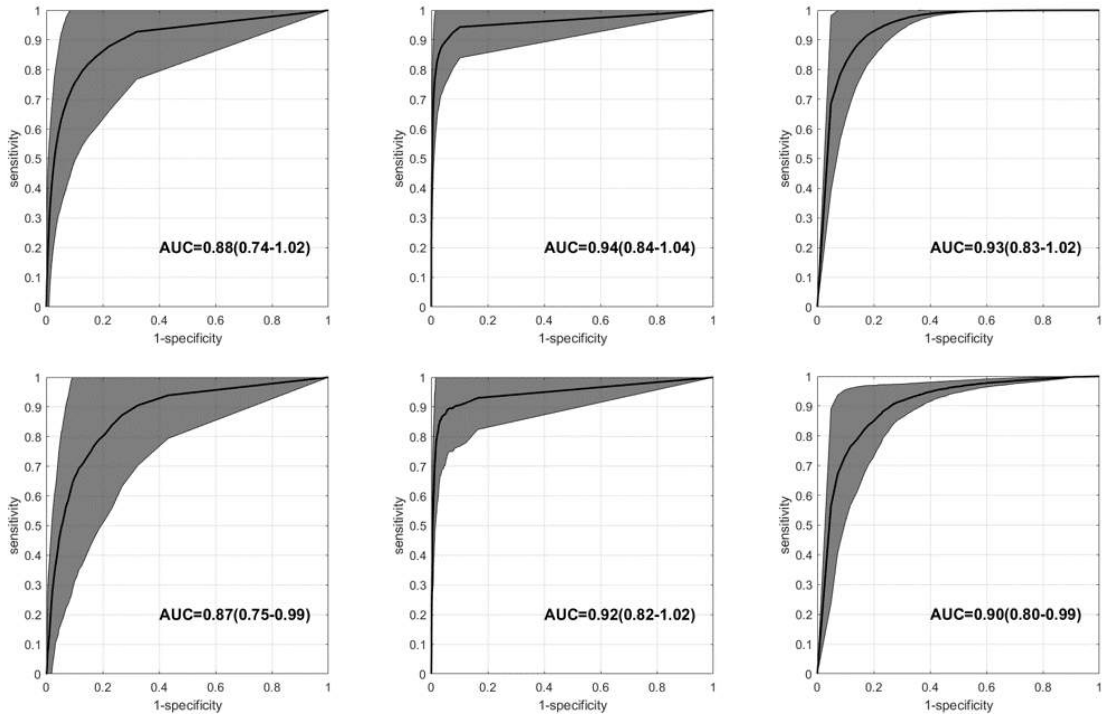


(a) Focal (b) Generalized (c) Normal

Figure C.8: Upper row: average ROC curves of the VGG model applied to the training set of Set G; bottom row: average ROC curves of the VGG model applied to the test set of Set G. These were built based on the results of 5-fold cross-validation. The 95% CI of the ROC curves is shown as a shaded area. The resulting AUC value and corresponding 95% CIs are also presented.

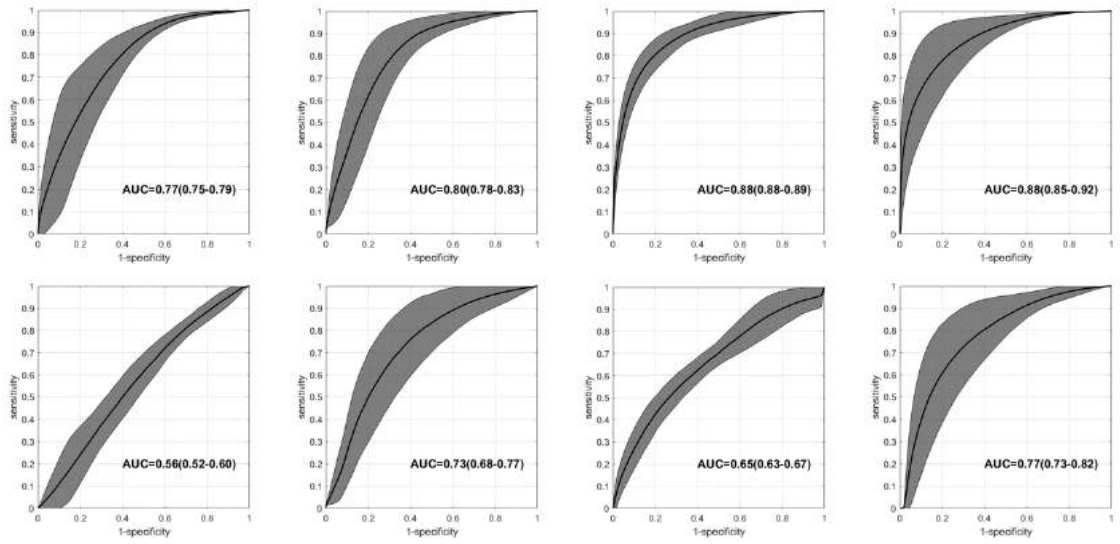
Table C.6: Average per class accuracy (Acc), sensitivity (Sens) and specificity (Spec) for the VGG model trained with Set G and Set H. These values were calculated on the test set, based on the results of 5-fold cross-validation, using a threshold where the sensitivity is equal to the specificity. The 95% CIs of each parameter are also presented.

	Set G			Set H		
Class	Acc (%)	Sens (%)	Spec (%)	Acc (%)	Sens (%)	Spec (%)
Focal	60.05 (51.01-69.09)	59.53 (32.05-87.01)	60.24 (38.98-81.50)	79.37 (63.73-95.01)	80.65 (51.76-100.00)	79.34 (62.88-95.80)
Generalized	56.06 (51.60-60.52)	57.29 (38.80-75.77)	55.44 (39.76-71.11)	89.99 (82.87-97.11)	90.61 (76.71-100.00)	89.99 (82.69-97.28)
Normal	95.11 (60.79-69.43)	64.42 (55.10-73.75)	65.56 (52.50-78.62)	82.34 (68.85-95.84)	82.32 (67.81-96.83)	83.13 (58.91-100.00)



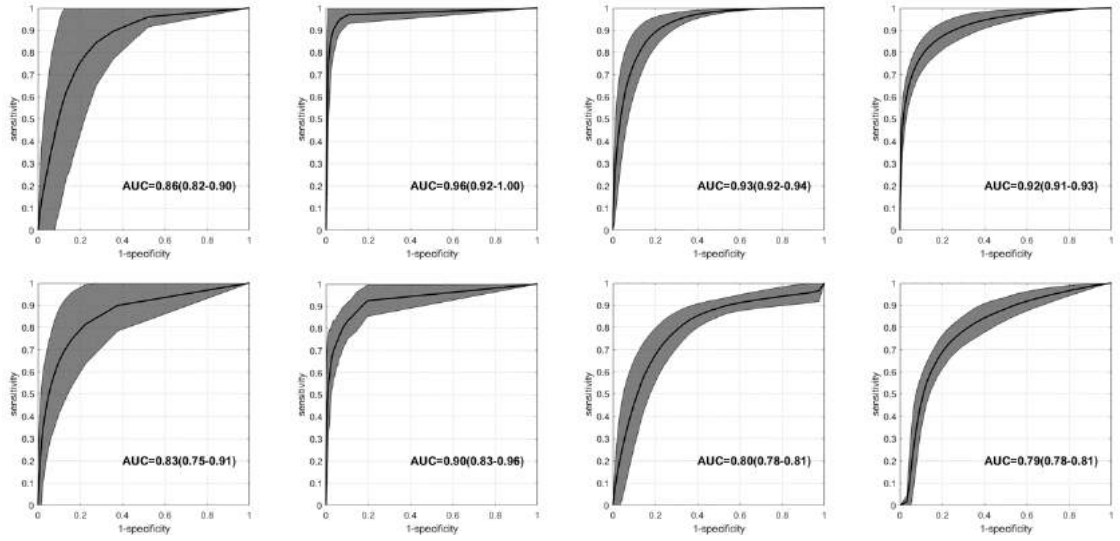
(a) Focal (b) Generalized (c) Normal

Figure C.9: Upper row: average ROC curves of the VGG model applied to the training set of Set H; bottom row: average ROC curves of the VGG model applied to the test set of Set H. These were built based on the results of 5-fold cross-validation. The 95% CI of the ROC curves is shown as a shaded area. The resulting AUC value and corresponding 95% CIs are also presented.



(a) Focal (b) Generalized (c) Normal (d) Abnormal

Figure C.10: Upper row: average ROC curves of the VGG model applied to the training set of Set I; bottom row: average ROC curves of the VGG model applied to the test set of Set I. These were built based on the results of 5-fold cross-validation. The 95% CI of the ROC curves is shown as a shaded area. The resulting AUC value and corresponding 95% CIs are also presented.



(a) Focal (b) Generalized (c) Normal (d) Abnormal

Figure C.11: Upper row: average ROC curves of the VGG model applied to the training set of Set J; bottom row: average ROC curves of the VGG model applied to the test set of Set J. These were built based on the results of 5-fold cross-validation. The 95% CI of the ROC curves is shown as a shaded area. The resulting AUC value and corresponding 95% CIs are also presented.

Table C.7: Average per class accuracy (Acc), sensitivity (Sens) and specificity (Spec) for the VGG model trained with Set I and Set J. These values were calculated on the test set, based on the results of 5-fold cross-validation, using a threshold where the sensitivity is equal to the specificity. The 95% CIs of each parameter are also presented.

	Set I			Set J		
Class	Acc (%)	Sens (%)	Spec (%)	Acc (%)	Sens (%)	Spec (%)
Focal	55.51 (47.42-63.61)	55.15 (45.93-64.37)	55.59 (44.37-66.81)	77.34 (59.00-95.67)	79.03 (51.76-100.00)	77.31 (58.36-96.26)
Generalized	67.87 (61.33-74.40)	67.69 (49.28-86.10)	67.91 (55.45-80.37)	85.54 (76.01-95.08)	87.52 (78.75-96.30)	85.53 (75.92-95.15)
Normal	61.11 (57.49-64.74)	61.66 (55.70-67.61)	60.72 (50.97-70.46)	74.37 (71.26-77.47)	74.20 (64.30-84.10)	74.65 (65.12-84.17)
Abnormal	71.17 (58.91-83.43)	71.40 (53.07-90.31)	71.11 (51.91-90.31)	74.76 (72.55-76.96)	74.49 (67.62-81.35)	74.90 (68.11-81.70)

References

[1] WO Tatum, G Rubboli, PW Kaplan, SM Mirsatari, D Gloss, LO K Caboclo, FW Drislane, M Koutroumanidis, DL Schomer, D Kastelijn-Nolst Trenite, et al. Clinical utility of eeg in diagnosing and monitoring epilepsy in adults. *Clinical Neurophysiology*, 2018.

[2] Fernando Lopes Da Silva, Wouter Blanes, Stiliyan N Kalitzin, Jaime Parra, Piotr Suffczynski, and Demetrios N Velis. Epilepsies as dynamical diseases of brain systems: basic models of the transition between normal and epileptic activity. *Epilepsia*, 44:72–83, 2003.

[3] Robert S Fisher, Carlos Acevedo, Alexis Arzimanoglou, Alicia Bogacz, J Helen Cross, Christian E Elger, Jerome Engel Jr, Lars Forsgren, Jacqueline A French, Mike Glynn, et al. Ilae official report: a practical clinical definition of epilepsy. *Epilepsia*, 55(4):475–482, 2014.

[4] Amir Zaidi, Peter Clough, Paul Cooper, Bruce Scheepers, and Adam P Fitzpatrick. Misdiagnosis of epilepsy: many seizure-like attacks have a cardiovascular cause. *Journal of the American College of Cardiology*, 36(1):181–184, 2000.

[5] R Beach and Richard Reading. The importance of acknowledging clinical uncertainty in the diagnosis of epilepsy and non-epileptic events. *Archives of disease in childhood*, 90(12):1219–1222, 2005.

[6] Fernando Torres. Atlas and classification of electroencephalography. *Pediatric Neurology*, 22(4):332, 2000.

[7] William J Nowack. Epilepsy: a costly misdiagnosis. *Clinical Electroencephalography*, 28(4):225–228, 1997.

[8] Soheyl Noachtar and Jan Rémi. The role of eeg in epilepsy: a critical review. *Epilepsy & Behavior*, 15(1):22–33, 2009.

[9] Thaddeus S Walczak, Rodney A Radtke, and Darrel V Lewis. Accuracy and interobserver reliability of scalp ictal eeg. *Neurology*, 42(12):2279–2279, 1992.

[10] Nancy Foldvary, G Klem, J Hammel, W Bingaman, I Najm, and H Lüders. The localizing value of ictal eeg in focal epilepsy. *Neurology*, 57(11):2022–2028, 2001.

[11] Jyoti Pillai and Michael R Sperling. Interictal eeg and the diagnosis of epilepsy. *Epilepsia*, 47:14–22, 2006.

[12] SJM Smith. Eeg in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 2):ii2–ii7, 2005.

[13] Shaun S Lodder, Jessica Askamp, and Michel JAM van Putten. Computer-assisted interpretation of the eeg background pattern: a clinical evaluation. *PloS one*, 9(1):e85966, 2014.

[14] Shaun S Lodder and Michel JAM van Putten. Automated eeg analysis: Characterizing the posterior dominant rhythm. *Journal of neuroscience methods*, 200(1):86–93, 2011.

[15] Shaun S Lodder and Michel JAM van Putten. Quantification of the adult eeg background pattern. *Clinical neurophysiology*, 124(2):228–237, 2013.

[16] Jeffrey W Britton, Lauren C Frey, JL Hopp, P Korb, MZ Koubeissi, WE Lievens, EM Pestana-Knight, and EK Louis St. *Electroencephalography (EEG): An introductory text and atlas of normal and abnormal findings in adults, children, and infants*. American Epilepsy Society, Chicago, 2016.

[17] J Craig Henry. Electroencephalography: Basic principles, clinical applications, and related fields. *Neurology*, 67(11):2092–2092, 2006.

[18] Donald L Schomer and Fernando Lopes Da Silva. *Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2012.

[19] Michal Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002.

[20] Michel Jam Van Putten. *Essentials of neurophysiology: basic concepts and clinical applications for scientists and engineers*. Springer Publishing Company, Incorporated, 2009.

[21] HVMN. Eeg recording. URL: https://hvmn.com/biohacker-guide/cognition/eeg-measures-of-cognition.

[22] Richard Caton. Electrical currents of the brain. *The Journal of Nervous and Mental Disease*, 2(4):610, 1875.

[23] Richard Caton. Iv.—interim report on investigation of the electric currents of the brain. *American Journal of EEG Technology*, 11(1):23–24, 1971.

[24] Hans Berger. Über das elektrenkephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten*, 87(1):527–570, 1929.

[25] David Millett. Hans berger: From psychic energy to the eeg. *Perspectives in biology and medicine*, 44(4):522–542, 2001.

[26] Edgar Douglas Adrian and Kazumi Yamagiwa. The origin of the berger rhythm. *Brain: A Journal of Neurology*, 1935.

[27] Edgar D Adrian and Bryan HC Matthews. The interpretation of potential waves in the cortex. *The Journal of Physiology*, 81(4):440–471, 1934.

[28] Pierre Gloor. Hans berger on electroencephalography. *American Journal of EEG Technology*, 9(1):1–8, 1969.

[29] Mary AB Brazier. Pioneers in the discovery of evoked potentials. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 59(1):2–8, 1984.

[30] Thomas F Collura. History and evolution of electroencephalographic instruments and techniques. *Journal of clinical neurophysiology*, 10(4):476–504, 1993.

[31] Frederic A Gibbs. Interpretation of the electroencephalogram. *The Journal of Psychology*, 4(2):365–382, 1937.

[32] Alexander Forbes and Catharine Thacher. Amplification of action currents with the electron tube in recording with the string galvanometer. *American Journal of Physiology-Legacy Content*, 52(3):409–471, 1920.

[33] EL Garceau and H Davis. An ink-writing electro-encephalograph. *Archives of Neurology & Psychiatry*, 34(6):1292–1294, 1935.

[34] E Lovett Garceau and Hallowell Davis. An amplifier, recording system, and stimulating devices for the study of cerebral action currents. *American Journal of Physiology-Legacy Content*, 107(2):305–310, 1934.

[35] Albert M Grass. The electroencephalographic heritage until 1960. *American Journal of EEG Technology*, 24(3):133–173, 1984.

[36] Richard W Homan, John Herman, and Phillip Purdy. Cerebral location of international 10–20 system electrode placement. *Electroencephalography and clinical neurophysiology*, 66(4):376–382, 1987.

[37] George H Klem, Hans Otto Lüders, HH Jasper, C Elger, et al. The ten-twenty electrode system of the international federation. *Electroencephalogr Clin Neurophysiol*, 52(3):3–6, 1999.

[38] Vernon L Towle, José Bolaños, Diane Suarez, Kim Tan, Robert Grzeszczuk, David N Levin, Raif Cakmur, Samuel A Frank, and Jean-Paul Spire. The spatial location of eeg electrodes: locating the best-fitting sphere relative to cortical anatomy. *Electroencephalography and clinical neurophysiology*, 86(1):1–6, 1993.

[39] Lin Zhu, Haifeng Chen, Xu Zhang, Kai Guo, Shujing Wang, Yu Wang, Weihua Pei, and Hongda Chen. Design of portable multi-channel eeg signal acquisition system. In *Biomedical Engineering and Informatics, 2009. BMEI'09. 2nd International Conference on*, pages 1–4. IEEE, 2009.

[40] Rui Martins, Siegfried Selberherr, and Francisco A Vaz. A cmos ic for portable eeg acquisition systems. *IEEE Transactions on Instrumentation and measurement*, 47(5):1191–1196, 1998.

[41] Marc R Nuwer, Giancarlo Comi, Ronald Emerson, Anders Fuglsang-Frederiksen, Jean-Michel Guérit, Hermann Hinrichs, Akio Ikeda, Fransisco Jose C Luccas, and Peter Rappelsburger. Ifcn standards for digital recording of clinical eeg. *Clinical Neurophysiology*, 106(3):259–261, 1998.

[42] Krikor Tufenkjian. Eeg instrumentation, montage, polarity, and localization. In *Epilepsy Board Review*, pages 15–32. Springer, 2017.

[43] Jürgen Kayser and Craig E Tenke. Issues and considerations for using the scalp surface laplacian in eeg/erp research: a tutorial review. *International Journal of Psychophysiology*, 97(3):189–209, 2015.

[44] J Gotman and P Gloor. Automatic recognition and quantification of interictal epileptic activity in the human scalp eeg. *Electroencephalography and clinical neurophysiology*, 41(5):513–529, 1976.

[45] AC Da Rosa, B Kemp, T Paiva, FH Lopes da Silva, and HAC Kamphuisen. A model-based detector of vertex waves and k complexes in sleep electroencephalogram. *Electroencephalography and clinical Neurophysiology*, 78(1):71–79, 1991.

[46] Shlomit Yuval-Greenberg, Orr Tomer, Alon S Keren, Israel Nelken, and Leon Y Deouell. Transient induced gamma-band response in eeg as a manifestation of miniature saccades. *Neuron*, 58(3):429–441, 2008.

[47] Gyorgy Buzsaki. *Rhythms of the Brain*. Oxford University Press, 2006.

[48] John S Ebersole and Timothy A Pedley. Current practice of clinical electroencephalography, 3rd edn. *European Journal of Neurology*, 10(5):604–605, 2003.

[49] Paul L Nunez and Ramesh Srinivasan. Electric fields of the brain: The neurophysics of eeg. *Physics Today*, 35(6):59, 1982.

[50] Masaki Iwasaki, Christoph Kellinghaus, Andreas V Alexopoulos, Richard C Burgess, Arun N Kumar, Yanning H Han, Hans O Lüders, and R John Leigh. Effects of eyelid closure, blinks, and eye movements on the electroencephalogram. *Clinical Neurophysiology*, 116(4):878–885, 2005.

[51] Otavio G Lins, Terence W Picton, Patrick Berg, and Michael Scherg. Ocular artifacts in eeg and event-related potentials i: Scalp topography. *Brain topography*, 6(1):51–63, 1993.

[52] Stefan Debener, Cornelia Kranczioch, and Ingmar Gutberlet. Eeg quality: origin and reduction of the eeg cardiac-related artefact. In *EEG-fMRI*, pages 135–151. Springer, 2009.

[53] Thomas C Ferree, Phan Luu, Gerald S Russell, and Don M Tucker. Scalp electrode impedance, infection risk, and eeg data quality. *Clinical Neurophysiology*, 112(3):536–544, 2001.

[54] Carrie A Joyce, Irina F Gorodnitsky, and Marta Kutas. Automatic removal of eye movement and blink artifacts from eeg data using blind component separation. *Psychophysiology*, 41(2):313–325, 2004.

[55] Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-Won Lee, Martin J Mckeown, Vicente Iragui, and Terrence J Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178, 2000.

[56] Hugh Nolan, Robert Whelan, and RB Reilly. Faster: fully automated statistical thresholding for eeg artifact rejection. *Journal of neuroscience methods*, 192(1):152–162, 2010.

[57] J Gotman, D Flanagan, J Zhang, and B Rosenblatt. Automatic seizure detection in the newborn: methods and initial evaluation. *Electroencephalography and clinical neurophysiology*, 103(3):356–362, 1997.

[58] Michael JAM van Putten. Nearest neighbor phase synchronization as a measure to detect seizure activity from scalp eeg recordings. *Journal of clinical neurophysiology*, 20(5):320–325, 2003.

[59] Michel JAM Van Putten, Taco Kind, Frank Visser, and Vera Lagerburg. Detecting temporal lobe seizures from scalp eeg recordings: a comparison of various features. *Clinical neurophysiology*, 116(10):2480–2489, 2005.

[60] Scott B Wilson and Ronald Emerson. Spike detection: a review and comparison of algorithms. *Clinical Neurophysiology*, 113(12):1873–1881, 2002.

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

[62] Michel JAM van Putten, Jeannette Hofmeijer, Barry J Ruijter, and Marleen C Tjepkema-Cloostermans. Deep learning for outcome prediction of postanoxic coma. In *EMBEC & NBC 2017*, pages 506–509. Springer, 2017.

[63] RA Segrave, RH Thomson, Nicholas R Cooper, RJ Croft, DM Sheppard, and PB Fitzgerald. Upper alpha activity during working memory processing reflects abnormal inhibition in major depression. *Journal of affective disorders*, 127(1-3):191–198, 2010.

[64] D Spronk, M Arns, KJ Barnett, NJ Cooper, and E Gordon. An investigation of eeg, genetic and cognitive markers of treatment response to antidepressant medication in patients with major depressive disorder: a pilot study. *Journal of affective disorders*, 128(1-2):41–48, 2011.

[65] Maria G Knyazeva, Mahdi Jalili, Reto Meuli, Martin Hasler, Oscar De Feo, and Kim Q Do. Alpha rhythm and hypofrontality in schizophrenia. *Acta Psychiatrica Scandinavica*, 118(3):188–199, 2008.

[66] Yi Jin, Steven G Potkin, Aaron S Kemp, Steven T Huerta, Gustavo Alva, Trung Minh Thai, Danilo Carreon, and William E Bunney Jr. Therapeutic effects of individualized alpha frequency transcranial magnetic stimulation (αtms) on the negative symptoms of schizophrenia. *Schizophrenia bulletin*, 32(3):556–561, 2005.

[67] Ryouhei Ishii, Leonides Canuet, Ryu Kurimoto, Koji Ikezawa, AOKI Yasunori, Michiyo Azechi, Hidetoshi Takahashi, Takayuki Nakahachi, Masao Iwase, Hiroaki Kazui, et al. Frontal shift of posterior alpha activity is correlated with cognitive impairment in early alzheimer's disease: A magnetoencephalography–beamformer study. *Psychogeriatrics*, 10(3):138–143, 2010.

[68] Seung-Hwan Lee, Young-Min Park, Do-Won Kim, and Chang-Hwan Im. Global synchronization index as a biological correlate of cognitive decline in alzheimer's disease. *Neuroscience research*, 66(4):333–339, 2010.

[69] Leonide Goldstein, Neal W Stoltzfus, and Joseph F Gardocki. Changes in interhemispheric amplitude relationships in the eeg during sleep. *Physiology & Behavior*, 8(5):811–815, 1972.

[70] William Dement and Nathaniel Kleitman. Cyclic variations in eeg during sleep and their relation to eye movements, body motility, and dreaming. *Electroencephalography and clinical neurophysiology*, 9(4):673–690, 1957.

[71] Leslie C Jameson and Tod B Sloan. Using eeg to monitor anesthesia drug effects during surgery. *Journal of clinical monitoring and computing*, 20(6):445–472, 2006.

[72] Ivan Cohen, Vincent Navarro, Stéphane Clemenceau, Michel Baulac, and Richard Miles. On the origin of interictal activity in human temporal lobe epilepsy in vitro. *Science*, 298(5597):1418–1421, 2002.

[73] A Hufnagel, M Dümpelmann, J Zentner, O Schijns, and CE Elger. Clinical relevance of quantified intracranial interictal spike activity in presurgical evaluation of epilepsy. *Epilepsia*, 41(4):467–478, 2000.

[74] Brian M Dale, Mark A Brown, and Richard C Semelka. *MRI: basic principles and applications*. John Wiley & Sons, 2015.

[75] Scott H Faro and Feroze B Mohamed. *Functional MRI: basic principles and clinical applications*. Springer Science & Business Media, 2006.

[76] Jiang Hsieh et al. Computed tomography: principles, design, artifacts, and recent advances. SPIE Bellingham, WA, 2009.

[77] Michael E Phelps. *PET: molecular imaging and its biological applications*. Springer Science & Business Media, 2004.

[78] H Stefan, G Pawlik, HG Böcher-Schwarz, HJ Biersack, W Burr, H Penin, and W-D Heiss. Functional and morphological abnormalities in temporal lobe epilepsy: a comparison of interictal and ictal eeg, ct, mri, spect and pet. *Journal of neurology*, 234(6):377–384, 1987.

[79] Conall J Garvey and Rebecca Hanlon. Computed tomography in clinical practice. *BMJ: British Medical Journal*, 324(7345):1077, 2002.

[80] Thomas R Browne and Gregory L Holmes. *Handbook of epilepsy*. Jones & Bartlett Learning, 2008.

[81] Mary Jane England, Catharyn T Liverman, Andrea M Schultz, and Larisa M Strawbridge. Epilepsy across the spectrum: Promoting health and understanding.: A summary of the institute of medicine report. *Epilepsy & Behavior*, 25(2):266–276, 2012.

[82] Hanneke M de Boer. Epilepsy stigma: moving from a global problem to global solutions. *Seizure*, 19(10):630–636, 2010.

[83] Robert A Gross. A brief history of epilepsy and its therapy in the western hemisphere. *Epilepsy research*, 12(2):65–74, 1992.

[84] James Longrigg. Epilepsy in ancient greek medicine—the vital step. *Seizure-European Journal of Epilepsy*, 9(1):12–21, 2000.

[85] Emmanouil Magiorkinis, Kalliopi Sidiropoulou, and Aristidis Diamantis. Hallmarks in the history of epilepsy: epilepsy in antiquity. *Epilepsy & Behavior*, 17(1):103–108, 2010.

[86] J Chadwick Hippocrates and WN Mann. *On the sacred disease*. DC Stevenson, Web Atomics, 2000.

[87] James W Wheless, James Willmore, and Roger A Brumback. *Advanced therapy in epilepsy*. PMPH-USA, 2009.

[88] Louis Jean François Delasiauve. *Traité de l'épilepsie*. Masson, 1854.

[89] John Russell Reynolds. *The diagnosis of diseases of the brain, spinal cord, nerves, and their appendages*. J. Churchill, 1855.

[90] WR Gowers. Epilepsy and other chronic convulsive disorders. *London: Churchill*, 223, 1881.

[91] Robert S Fisher, Walter Van Emde Boas, Warren Blume, Christian Elger, Pierre Genton, Phillip Lee, and Jerome Engel Jr. 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, 2005.

[92] Gregory L Holmes, Steven C Schachter, and Dorothee GA Kasteleijn-Nolst Trenite. *Behavioral aspects of epilepsy: principles and practice*. Demos Medical Publishing, 2007.

[93] John S Duncan, Josemir W Sander, Sanjay M Sisodiya, and Matthew C Walker. Adult epilepsy. *The Lancet*, 367(9516):1087–1100, 2006.

[94] William G Lennox. The heredity of epilepsy as told by relatives and twins. *Journal of the American medical association*, 146(6):529–536, 1951.

[95] Jerome Engel. A proposed diagnostic scheme for people with epileptic seizures and with epilepsy: report of the ilae task force on classification and terminology. *Epilepsia*, 42(6):796–803, 2001.

[96] John F Annegers, Walter A Rocca, and W Allen Hauser. Causes of epilepsy: contributions of the rochester epidemiology project. In *Mayo Clinic Proceedings*, volume 71, pages 570–575. Elsevier, 1996.

[97] Samuel F Berkovic, John C Mulley, Ingrid E Scheffer, and Steven Petrou. Human epilepsies: interaction of genetic and acquired factors. *Trends in neurosciences*, 29(7):391–397, 2006.

[98] Irene Kotsopoulos, Marc de Krom, Fons Kessels, Jan Lodder, Jaap Troost, Mascha Twellaar, Tiny van Merode, and André Knottnerus. Incidence of epilepsy and predictive factors of epileptic and non-epileptic seizures. *Seizure*, 14(3):175–182, 2005.

[99] Devender Bhalla, Bertrand Godet, Michel Druet-Cabanac, and Pierre-Marie Preux. Etiologies of epilepsy: a comprehensive review. *Expert review of neurotherapeutics*, 11(6):861–876, 2011.

[100] Anne T Berg. Risk of recurrence after a first unprovoked seizure. *Epilepsia*, 49:13–18, 2008.

[101] Jerome Engel. *Seizures and epilepsy*, volume 83. Oxford University Press, 2013.

[102] Ethan M Goldberg and Douglas A Coulter. Mechanisms of epileptogenesis: a convergence on neural circuit dysfunction. *Nature Reviews Neuroscience*, 14(5):337, 2013.

[103] David Spencer. Auras are frequent in patients with generalized epilepsy. *Epilepsy currents*, 15(2):75–77, 2015.

[104] AK Gupta, PM Jeavons, RC Hughes, and A Covanis. Aura in temporal lobe epilepsy: clinical and electroencephalographic correlation. *Journal of Neurology, Neurosurgery & Psychiatry*, 46(12):1079–1083, 1983.

[105] Dileep R Nair, Imad Najm, Juan Bulacio, and Hans Lüders. Painful auras in focal epilepsy. *Neurology*, 57(4):700–702, 2001.

[106] David C Taylor and Moira Lochery. Temporal lobe epilepsy: origin and significance of simple and complex auras. *Journal of Neurology, Neurosurgery & Psychiatry*, 50(6):673–681, 1987.

[107] Maromi Nei and Ritu Bagla. Seizure-related injury and death. *Current neurology and neuroscience reports*, 7(4):335–341, 2007.

[108] Anna L Devlin, Morris Odell, Judith L Charlton, and Sjaanie Koppel. Epilepsy and driving: Current status of research. *Epilepsy research*, 102(3):135–152, 2012.

[109] Hsiu-Fang Chen, Yun-Fang Tsai, Yea-Pyng Lin, Mo-Song Shih, and Jui-Chen Chen. The relationships among medicine symptom distress, self-efficacy, patient–provider relationship, and medication compliance in patients with epilepsy. *Epilepsy & Behavior*, 19(1):43–49, 2010.

[110] Marcelo E Lancman, William J Craven, Jorge J Asconapé, and J Kiffin Penry. Clinical management of recurrent postictal psychosis. *Journal of Epilepsy*, 7(1):47–51, 1994.

[111] Robert S Fisher and Steven C Schachter. The postictal state: a neglected entity in the management of epilepsy. *Epilepsy & Behavior*, 1(1):52–59, 2000.

[112] Fawaz Al-Mufti and Jan Claassen. Neurocritical care: status epilepticus review. *Critical care clinics*, 30(4):751–764, 2014.

[113] WA Hauser. Status epilepticus: frequency, etiology, and neurological sequelae. *Advances in neurology*, 34:3–14, 1983.

[114] Daniel H Lowenstein and Brian K Alldredge. Status epilepticus. *New England Journal of Medicine*, 338(14):970–976, 1998.

[115] Peter Shearer, David Park, Andrew J Bowman, and Stephen J Huff. Seizures and status epilepticus: Diagnosis and management in the emergency department. *Emergency Medicine Practice+ Em Practice Guidelines Update*, 8(8):1–31, 2006.

[116] Nikolas Hitiris, Rajiv Mohanraj, John Norrie, and Martin J Brodie. Mortality in epilepsy. *Epilepsy & Behavior*, 10(3):363–376, 2007.

[117] Gretchen L Birbeck, Ron D Hays, Xinping Cui, and Barbara G Vickrey. Seizure reduction and quality of life improvements in people with epilepsy. *Epilepsia*, 43(5):535–538, 2002.

[118] FA Chowdhury, L Nashef, and RDC Elwes. Misdiagnosis in epilepsy: a review and recognition of diagnostic uncertainty. *European journal of neurology*, 15(10):1034–1042, 2008.

[119] Shufang Li, Weidong Zhou, Qi Yuan, Shujuan Geng, and Dongmei Cai. Feature extraction and recognition of ictal eeg using emd and svm. *Computers in biology and medicine*, 43(7):807–816, 2013.

[120] Marco de Curtis, John GR Jefferys, and Massimo Avoli. Interictal epileptiform discharges in partial epilepsy. 2012.

[121] Allan Krumholz. Epilepsy: a comprehensive textbook. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 45(6):825–825, 1999.

[122] Simon Shorvon, Emilio Perucca, and Jerome Engel Jr. *The treatment of epilepsy*. John Wiley & Sons, 2015.

[123] Clare M Eddy, Hugh E Rickards, and Andrea E Cavanna. The cognitive impact of antiepileptic drugs. *Therapeutic advances in neurological disorders*, 4(6):385–407, 2011.

[124] Mervyn J Eadie. Shortcomings in the current treatment of epilepsy. *Expert review of neurotherapeutics*, 12(12):1419–1427, 2012.

[125] Morris Sabin and Harry Oxorn. Epilepsy and pregnancy. *Obstetrics & Gynecology*, 7(2):175–179, 1956.

[126] Louis St and K Erik. Minimizing aed adverse effects: improving quality of life in the interictal state in epilepsy care. *Current neuropharmacology*, 7(2):106–114, 2009.

[127] Piero Perucca and Frank G Gilliam. Adverse effects of antiepileptic drugs. *The Lancet Neurology*, 11(9):792–802, 2012.

[128] Gregory K Bergey. Neurostimulation in the treatment of epilepsy. *Experimental neurology*, 244:87–95, 2013.

[129] Barbara C Jobst and Gregory D Cascino. Resective epilepsy surgery for drug-resistant focal epilepsy: a review. *Jama*, 313(3):285–293, 2015.

[130] J Millichap. Epilepsy surgery outcome. *Pediatric Neurology Briefs*, 12(12), 1998.

[131] Elinor Ben-Menachem. Vagus-nerve stimulation for the treatment of epilepsy. *The Lancet Neurology*, 1(8):477–482, 2002.

[132] Alan Guberman. Vagus nerve stimulation in the treatment of epilepsy. *Canadian Medical Association Journal*, 171(10):1165–1166, 2004.

[133] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[134] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2009.

[135] Pat Langley and Jaime G Carbonell. Approaches to machine learning. *Journal of the American Society for Information Science*, 35(5):306–316, 1984.

[136] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[137] David M Dutton and Gerard V Conroy. A review of machine learning. *The knowledge engineering review*, 12(4):341–367, 1997.

[138] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[139] Haohan Wang, Bhiksha Raj, and Eric P Xing. On the origin of deep learning. *arXiv preprint arXiv:1702.07800*, 2017.

[140] Jaime G Carbonell, Ryszard S Michalski, and Tom M Mitchell. An overview of machine learning. In *Machine Learning, Volume I*, pages 3–23. Elsevier, 1983.

[141] Pamela McCorduck. *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. AK Peters/CRC Press, 2009.

[142] David L Poole and Alan K Mackworth. *Artificial Intelligence: foundations of computational agents*. Cambridge University Press, 2010.

[143] Pat Langley. *Elements of machine learning*. Morgan Kaufmann, 1996.

[144] Pat Langley. The changing science of machine learning. *Machine Learning*, 82(3):275–279, 2011.

[145] John MacCormick. *Nine algorithms that changed the future: The ingenious ideas that drive today's computers*. Princeton University Press, 2011.

[146] Nicolas Rashevsky. *Mathematical Biophysics. Rev*. The University Of Chicago; Chicago, Illinois, 1948.

[147] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[148] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[149] Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, CORNELL AERONAUTICAL LAB INC BUFFALO NY, 1961.

[150] KC Fu. *Sequential methods in pattern recognition and machine learning*, volume 52. Academic press, 1968.

[151] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[152] Nils J Nilsson. Learning machines. 1965.

[153] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[154] Patrick H Winston. Learning and reasoning by analogy. *Communications of the ACM*, 23(12):689–703, 1980.

[155] Earl B Hunt, Janet Marin, and Philip J Stone. Experiments in induction. 1966.

[156] Bruce G Buchanan and Tom M Mitchell. Model-directed learning of production rules. In *Pattern-directed inference systems*, pages 297–312. Elsevier, 1978.

[157] Marvin Minsky and Seymour Papert. Perceptrons. 1969.

[158] Sir James Lighthill. Artificial intelligence: A general survey. part i of artificial intelligence': a paper symposium. london: Science research council, 1973.

[159] National Research Council et al. *Funding a revolution: Government support for computing research*. National Academies Press, 1999.

[160] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[161] Hans-Dieter Block. The perceptron: A model for brain functioning. i. *Reviews of Modern Physics*, 34(1):123, 1962.

[162] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[163] Geoffrey E Hinton. How neural networks learn from experience. *Scientific American*, 267(3):144–151, 1992.

[164] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[165] Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

[166] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[167] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868, 2009.

[168] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*, 2014.

[169] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[170] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[171] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.

[172] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[173] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[174] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[175] Huimin Lu, Yujie Li, Min Chen, Hyoungseop Kim, and Seiichi Serikawa. Brain intelligence: go beyond artificial intelligence. *Mobile Networks and Applications*, 23(2):368–375, 2018.

[176] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[177] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

[178] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[179] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[180] Elizabeth Gibney. Google ai algorithm masters ancient game of go. *Nature News*, 529(7587):445, 2016.

[181] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[182] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[183] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[184] Stephen Marsland. *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2011.

[185] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[186] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[187] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[188] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[189] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[190] Quoc V Le, Jiquan Ngiam, Adam Coates, Abhik Lahiri, Bobby Prochnow, and Andrew Y Ng. On optimization methods for deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 265–272. Omnipress, 2011.

[191] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[192] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[193] Ronald J Williams and David Zipser. Gradient-based learning algorithms for recurrent networks and their computational complexity. *Backpropagation: Theory, architectures, and applications*, 1:433–486, 1995.

[194] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.

[195] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2017.

[196] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

[197] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[198] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[199] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[200] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[201] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[202] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[203] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[204] Zhuwei Qin, Funxun Yu, Chenchen Liu, and Xiang Chen. How convolutional neural network see the world-a survey of convolutional neural network visualization methods. *arXiv preprint arXiv:1804.11191*, 2018.

[205] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016.

[206] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[207] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[208] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[209] Quoc V Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE, 2013.

[210] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[211] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[212] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. *Google Research Blog. Retrieved June*, 20(14):5, 2015.

[213] Donglai Wei, Bolei Zhou, Antonio Torrabla, and William Freeman. Understanding intra-class knowledge inside cnn. *arXiv preprint arXiv:1507.02379*, 2015.

[214] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. 2010.

[215] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2018–2025. IEEE, 2011.

[216] Emmanuel d'Angelo, Alexandre Alahi, and Pierre Vandergheynst. Beyond bits: Reconstructing images from local binary descriptors. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 935–938. IEEE, 2012.

[217] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

[218] Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1546, 2015.

[219] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4829–4837, 2016.

[220] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *arXiv preprint arXiv:1704.05796*, 2017.

[221] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. *arXiv preprint arXiv:1801.03454*, 2018.

[222] William H Crown. Potential application of machine learning in health outcomes research and some statistical cautions. *Value in Health*, 18(2):137–140, 2015.

[223] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 2017.

[224] Daniele Ravı, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2017.

[225] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: a review. *Computer methods and programs in biomedicine*, 2018.

[226] Polina Mamoshina, Armando Vieira, Evgeny Putin, and Alex Zhavoronkov. Applications of deep learning in biomedicine. *Molecular pharmaceutics*, 13(5):1445–1454, 2016.

[227] Florian Mormann, Ralph G Andrzejak, Christian E Elger, and Klaus Lehnertz. Seizure prediction: the long and winding road. *Brain*, 130(2):314–333, 2006.

[228] Alexandros T Tzallas, Markos G Tsipouras, Dimitrios G Tsalikakis, Evaggelos C Karvounis, Loukas Astrakas, Spiros Konitsiotis, and Margaret Tzaphlidou. Automated epileptic seizure detection methods: a review study. In *Epilepsy-histological, electroencephalographic and psychological aspects*. InTech, 2012.

[229] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals. *Computers in biology and medicine*, 100:270–278, 2018.

[230] Sriram Ramgopal, Sigride Thome-Souza, Michele Jackson, Navah Ester Kadish, Iván Sánchez Fernández, Jacquelyn Klehm, William Bosl, Claus Reinsberger, Steven Schachter, and Tobias Loddenkemper. Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. *Epilepsy & behavior*, 37:291–307, 2014.

[231] Jean Gotman. Automatic recognition of epileptic seizures in the eeg. *Electroencephalography and clinical Neurophysiology*, 54(5):530–540, 1982.

[232] Jean Gotman. Automatic seizure detection: improvements and evaluation. *Electroencephalography and clinical Neurophysiology*, 76(4):317–324, 1990.

[233] WRS Webber, Ronald P Lesser, Russell T Richardson, and Kerry Wilson. An approach to seizure detection using an artificial neural network (ann). *Electroencephalography and clinical Neurophysiology*, 98(4):250–272, 1996.

[234] AJ Gabor, RR Leach, and FU Dowla. Automated seizure detection using a self-organizing neural network. *Electroencephalography and clinical Neurophysiology*, 99(3):257–266, 1996.

[235] U Rajendra Acharya, S Vinitha Sree, Ang Peng Chuan Alvin, and Jasjit S Suri. Use of principal component analysis for automatic classification of epileptic eeg activities in wavelet framework. *Expert Systems with Applications*, 39(10):9072–9078, 2012.

[236] Abdulhamit Subasi and M Ismail Gursoy. Eeg signal classification using pca, ica, lda and support vector machines. *Expert systems with applications*, 37(12):8659–8666, 2010.

[237] Pari Jahankhani, Vassilis Kodogiannis, and Kenneth Revett. Eeg signal classification using wavelet feature extraction and neural networks. In *Modern Computing, 2006. JVA'06. IEEE John Vincent Atanasoff 2006 International Symposium on*, pages 120–124. IEEE, 2006.

[238] Abdulhamit Subasi. Eeg signal classification using wavelet feature extraction and a mixture of expert model. *Expert Systems with Applications*, 32(4):1084–1093, 2007.

[239] Hasan Ocak. Automatic detection of epileptic seizures in eeg using discrete wavelet transform and approximate entropy. *Expert Systems with Applications*, 36(2):2027–2036, 2009.

[240] A Temko, E Thomas, W Marnane, G Lightbody, and G Boylan. Eeg-based neonatal seizure detection with support vector machines. *Clinical Neurophysiology*, 122(3):464–473, 2011.

[241] Kemal Polat and Salih Güneş. Classification of epileptiform eeg using a hybrid system based on decision tree classifier and fast fourier transform. *Applied Mathematics and Computation*, 187(2):1017–1026, 2007.

[242] Samanwoy Ghosh-Dastidar, Hojjat Adeli, and Nahid Dadmehr. Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection. *IEEE Transactions on Biomedical Engineering*, 55(2):512–518, 2008.

[243] U Rajendra Acharya, S Vinitha Sree, G Swapna, Roshan Joy Martis, and Jasjit S Suri. Automated eeg analysis of epilepsy: a review. *Knowledge-Based Systems*, 45:147–165, 2013.

[244] N Kannathal, Min Lim Choo, U Rajendra Acharya, and PK Sadasivan. Entropies for detection of epilepsy in eeg. *Computer methods and programs in biomedicine*, 80(3):187–194, 2005.

[245] Ling Guo, Daniel Rivero, Julián Dorado, Cristian R Munteanu, and Alejandro Pazos. Automatic feature extraction using genetic programming: An application to epileptic eeg classification. *Expert Systems with Applications*, 38(8):10425–10436, 2011.

[246] Roshan Joy Martis, U Rajendra Acharya, Jen Hong Tan, Andrea Petznick, Ratna Yanti, Chua Kuang Chua, EY Kwee Ng, and Louis Tong. Application of empirical mode decomposition (emd) for automated detection of epilepsy using eeg signals. *International journal of neural systems*, 22(06):1250027, 2012.

[247] Samanwoy Ghosh-Dastidar and Hojjat Adeli. Improved spiking neural networks for eeg classification and epilepsy and seizure detection. *Integrated Computer-Aided Engineering*, 14(3):187–212, 2007.

[248] Samanwoy Ghosh-Dastidar and Hojjat Adeli. A new supervised learning algorithm for multiple spiking neural networks with application in epilepsy and seizure detection. *Neural networks*, 22(10):1419–1431, 2009.

[249] U Rajendra Acharya, S Vinitha Sree, Peng Chuan Alvin Ang, Ratna Yanti, and Jasjit S Suri. Application of non-linear and wavelet based features for the automated identification of epileptic eeg signals. *International journal of neural systems*, 22(02):1250002, 2012.

[250] Fernando H Lopes da Silva, Wouter Blanes, Stiliyan N Kalitzin, Jaime Parra, Piotr Suffczynski, and Demetrios N Velis. Dynamical diseases of brain systems: different routes to epileptic seizures. *IEEE Transactions on Biomedical Engineering*, 50(5):540–548, 2003.

[251] SS Viglione and GO Walsh. Proceedings: Epileptic seizure prediction. *Electroencephalography and clinical neurophysiology*, 39(4):435–436, 1975.

[252] Kais Gadhoumi, Jean-Marc Lina, Florian Mormann, and Jean Gotman. Seizure prediction for therapeutic devices: A review. *Journal of neuroscience methods*, 260:270–282, 2016.

[253] Paul R Carney, Stephen Myers, and James D Geyer. Seizure prediction: methods. *Epilepsy & behavior*, 22:S94–S101, 2011.

[254] Florian Mormann, Thomas Kreuz, Christoph Rieke, Ralph G Andrzejak, Alexander Kraskov, Peter David, Christian E Elger, and Klaus Lehnertz. On the predictability of epileptic seizures. *Clinical neurophysiology*, 116(3):569–587, 2005.

[255] Ardalan Aarabi and Bin He. A rule-based seizure prediction method for focal neocortical epilepsy. *Clinical Neurophysiology*, 123(6):1111–1122, 2012.

[256] Ardalan Aarabi and Bin He. Seizure prediction in hippocampal and neocortical epilepsy using a model-based approach. *Clinical Neurophysiology*, 125(5):930–940, 2014.

[257] Mojtaba Bandarabadi, César A Teixeira, Jalil Rasekhi, and António Dourado. Epileptic seizure prediction using relative spectral power features. *Clinical Neurophysiology*, 126(2):237–248, 2015.

[258] Piotr Mirowski, Deepak Madhavan, Yann LeCun, and Ruben Kuzniecky. Classification of patterns of eeg synchronization for seizure prediction. *Clinical neurophysiology*, 120(11):1927–1940, 2009.

[259] Arthur Petrosian, Danil Prokhorov, Richard Homan, Richard Dasheiff, and Donald Wunsch II. Recurrent neural network based prediction of epileptic seizures in intra-and extracranial eeg. *Neurocomputing*, 30(1-4):201–218, 2000.

[260] Martha Morrell. Brain stimulation for epilepsy: can scheduled or responsive neurostimulation stop seizures? *Current opinion in neurology*, 19(2):164–168, 2006.

[261] Arthur Guez, Robert D Vincent, Massimo Avoli, and Joelle Pineau. Adaptive treatment of epilepsy via batch-mode reinforcement learning. In *AAAI*, pages 1671–1678, 2008.

[262] Joelle Pineau, Arthur Guez, Robert Vincent, Gabriella Panuccio, and Massimo Avoli. Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach. *International journal of neural systems*, 19(04):227–240, 2009.

[263] Jing Zhou. A retrospective glance at automatic detection of epileptic spike in electroencephalogram. *European Journal of BioMedical Research*, 1(4):9–17, 2015.

[264] J Gotman and LY Wang. State-dependent spike detection: concepts and preliminary results. *Electroencephalography and clinical Neurophysiology*, 79(1):11–19, 1991.

[265] Jean Gotman and Li-Yan Wang. State dependent spike detection: validation. *Clinical Neurophysiology*, 83(1):12–18, 1992.

[266] Pedro Guedes De Oliveira, Carlos Queiroz, and Fernando Lopes Da Silva. Spike detection based on a pattern recognition approach using a microcomputer. *Electroencephalography and clinical neurophysiology*, 56(1):97–103, 1983.

[267] Wayne E Hostetler, Herbert J Doller, and Richard W Homan. Assessment of a computer program to detect epileptiform spikes. *Electroencephalography and clinical neurophysiology*, 83(1):1–11, 1992.

[268] Alison A Dingle, Richard D Jones, Grant J Carroll, and W Richard Fright. A multistage system to detect epileptiform activity in the eeg. *IEEE Transactions on Biomedical Engineering*, 40(12):1260–1268, 1993.

[269] T Pietilä, S Vapaakoski, U Nousiainen, A Värri, Hl Frey, V Häkkinen, and Y Neuvo. Evaluation of a computerized system for recognition of epileptic activity during long-term eeg recording. *Electroencephalography and clinical neurophysiology*, 90(6):438–443, 1994.

[270] R Benlamri, M Batouche, S Rami, and C Bouanaka. An automated system for analysis and interpretation of epileptiform activity in the eeg. *Computers in biology and medicine*, 27(2):129–139, 1997.

[271] Michael A Black, Richard D Jones, Grant J Carroll, Alison A Dingle, Ivan M Donaldson, and Philip J Parkin. Real-time detection of epileptiform activity in the eeg: a blinded clinical trial. *Clinical Electroencephalography*, 31(3):122–130, 2000.

[272] R Sankar and J Natour. Automatic computer analysis of transients in eeg. *Computers in biology and medicine*, 22(6):407–422, 1992.

[273] Mahmoud El-Gohary, James McNames, and Siegward Elsas. User-guided interictal spike detection. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 821–824. IEEE, 2008.

[274] K Vijayalakshmi and Appaji M Abhishek. Spike detection in epileptic patients eeg data using template matching technique. *International Journal of Computer Applications*, 2(6):5–8, 2010.

[275] Zhanfeng Ji, Takenao Sugi, Satoru Goto, Xingyu Wang, and Masatoshi Nakamura. Multi-channel template extraction for automatic eeg spike detection. In *Complex Medical Engineering (CME), 2011 IEEE/ICME International Conference on*, pages 179–184. IEEE, 2011.

[276] Zhanfeng Ji, Xingyu Wang, Takenao Sugi, Satoru Goto, and Masatoshi Nakamura. Automatic spike detection based on real-time multi-channel template. In *Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on*, volume 2, pages 648–652. IEEE, 2011.

[277] Antoine Nonclercq, Martine Foulon, Denis Verheulpen, Cathy De Cock, Marga Buzatu, Pierre Mathys, and Patrick Van Bogaert. Cluster-based spike detection algorithm adapts to interpatient and intrapatient variation in spike morphology. *Journal of neuroscience methods*, 210(2):259–265, 2012.

[278] Shaun S Lodder and Michel JAM van Putten. A self-adapting system for the automated detection of inter-ictal epileptiform discharges. *PloS one*, 9(1):e85180, 2014.

[279] Yung-Chun Liu, Chou-Ching K Lin, Jing-Jane Tsai, and Yung-Nien Sun. Model-based spike detection of epileptic eeg data. *Sensors*, 13(9):12536–12547, 2013.

[280] J Jing, J Dauwels, T Rakthanmanon, E Keogh, SS Cash, and MB Westover. Rapid annotation of interictal epileptiform discharges via template matching under dynamic time warping. *Journal of neuroscience methods*, 274:179–190, 2016.

[281] Shaun S Lodder, Jessica Askamp, and Michel JAM van Putten. Inter-ictal spike detection using a database of smart templates. *Clinical neurophysiology*, 124(12):2328–2335, 2013.

[282] WRS Webber, Brian Litt, K Wilson, and RP Lesser. Practical detection of epileptiform discharges (eds) in the eeg using an artificial neural network: a comparison of raw and parameterized eeg data. *Electroencephalography and clinical Neurophysiology*, 91(3):194–204, 1994.

[283] Christopher J James, Richard D Jones, Philip J Bones, and Grant J Carroll. Detection of epileptiform discharges in the eeg by a hybrid system comprising mimetic, self-organized artificial neural network, and fuzzy logic stages. *Clinical Neurophysiology*, 110(12):2049–2063, 1999.

[284] Scott B Wilson, Christine A Turner, Ronald G Emerson, and Mark L Scheuer. Spike detection ii: automatic, perception-based detection and clustering. *Clinical neurophysiology*, 110(3):404–411, 1999.

[285] C Castellaro, G Favaro, A Castellaro, A Casagrande, S Castellaro, DV Puthenparampil, and C Fattorello Salimbeni. An artificial intelligence approach to classify and analyse eeg traces. *Neurophysiologie Clinique/Clinical Neurophysiology*, 32(3):193–214, 2002.

[286] Hyun S Park, Yong H Lee, Nam G Kim, Doo-Soo Lee, and Sun I Kim. Detection of epileptiform activities in the eeg using neural network and expert system. *Studies in health technology and informatics*, 52:1255–1259, 1998.

[287] He Sheng Liu, Tong Zhang, and Fu Sheng Yang. A multistage, multimethod approach for automatic detection and classification of epileptiform eeg. *IEEE Transactions on biomedical engineering*, 49(12):1557–1566, 2002.

[288] Inan Güler and Elif Derya Übeyli. Adaptive neuro-fuzzy inference system for classification of eeg signals using wavelet coefficients. *Journal of neuroscience methods*, 148(2):113–121, 2005.

[289] Inan Guler and Elif Derya Ubeyli. Multiclass support vector machines for eeg-signals classification. *IEEE Transactions on Information Technology in Biomedicine*, 11(2):117–126, 2007.

[290] Elif Derya Übeyli. Analysis of eeg signals by combining eigenvector methods and multiclass support vector machines. *Computers in Biology and Medicine*, 38(1):14–22, 2008.

[291] Elif Derya Übeyli. Analysis of eeg signals by implementing eigenvector methods/recurrent neural networks. *Digital Signal Processing*, 19(1):134–143, 2009.

[292] Berdakh Abibullaev, Hee Don Seo, and Min Soo Kim. Epileptic spike detection using continuous wavelet transforms and artificial neural networks. *International journal of wavelets, multiresolution and information processing*, 8(01):33–48, 2010.

[293] Zainab Haydari, Yanqing Zhang, and Hamid Soltanian-Zadeh. Semi-automatic epilepsy spike detection from eeg signal using genetic algorithm and wavelet transform. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, pages 635–638. IEEE, 2011.

[294] Deng Wang, Duoqian Miao, and Chen Xie. Best basis-based wavelet packet entropy feature extraction and hierarchical eeg classification for epileptic detection. *Expert Systems with Applications*, 38(11):14314–14320, 2011.

[295] Umut Orhan, Mahmut Hekim, and Mahmut Ozer. Eeg signals classification using the k-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications*, 38(10):13475–13481, 2011.

[296] Patcharin Artameeyanant, Werapon Chiracharit, and Kosin Chamnongthai. Spike and epileptic seizure detection using wavelet packet transform based on approximate entropy and energy with artificial neural network. In *Biomedical Engineering International Conference (BMEiCON), 2012*, pages 1–5. IEEE, 2012.

[297] Esma Sezer, Hakan Işik, and Esra Saracoğlu. Employment and comparison of different artificial neural networks for epilepsy diagnosis from eeg signals. *Journal of Medical Systems*, 36(1):347–362, 2012.

[298] HN Suresh and Vinay Balasubramanyam. Wavelet transforms and neural network approach for epileptical eeg. In *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, pages 12–17. IEEE, 2013.

[299] Jonathan J Halford, Robert J Schalkoff, Jing Zhou, Selim R Benbadis, William O Tatum, Robert P Turner, Saurabh R Sinha, Nathan B Fountain, Amir Arain, Paul B Pritchard, et al. Standardized database development for eeg epileptiform transient detection: Eegnet scoring system and machine learning analysis. *Journal of neuroscience methods*, 212(2):308–316, 2013.

[300] Yuedong Song and Jiaxiang Zhang. Automatic recognition of epileptic eeg patterns via extreme learning machine and multiresolution feature extraction. *Expert Systems with Applications*, 40(14):5477–5489, 2013.

[301] Martha Feucht, Katrin Hoffmann, Karl Steinberger, Herbert Witte, Franz Benninger, Matthias Arnold, and Axel Doering. Simultaneous spike detection and topographic classification in pediatric surface eegs. *NeuroReport*, 8(9):2193–2197, 1997.

[302] Hojjat Adeli, Samanwoy Ghosh-Dastidar, and Nahid Dadmehr. A wavelet-chaos methodology for analysis of eegs and eeg subbands to detect seizure and epilepsy. *IEEE Transactions on Biomedical Engineering*, 54(2):205–211, 2007.

[303] Fernanda IM Argoud, Fernando M De Azevedo, José Marino Neto, and Eugênio Grillo. Sade 3: an effective system for automated detection of epileptiform events in long-term eeg based on context information. *Medical and Biological Engineering and Computing*, 44(6):459–470, 2006.

[304] Alexandros T Tzallas, Vaggelis P Oikonomou, and Dimitrios I Fotiadis. Epileptic spike detection using a kalman filter based approach. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 501–504. IEEE, 2006.

[305] Clement CC Pang, Adrian RM Upton, Glenn Shine, and Markad V Kamath. A comparison of algorithms for detection of spikes in the electroencephalogram. *IEEE Transactions on Biomedical Engineering*, 50(4):521–526, 2003.

[306] Tulga Kalayci and Ozcan Ozdamar. Wavelet preprocessing for automated neural network detection of eeg spikes. *IEEE engineering in medicine and biology magazine*, 14(2):160–166, 1995.

[307] Ö Özdamar and T Kalayci. Detection of spikes with artificial neural networks using raw eeg. *Computers and Biomedical Research*, 31(2):122–142, 1998.

[308] L Tarassenko, YU Khan, and MRG Holt. Identification of inter-ictal spikes in the eeg using neural network analysis. *IEE Proceedings-Science, Measurement and Technology*, 145(6):270–278, 1998.

[309] C Kurth, F Gilliam, and BJ Steinhoff. Eeg spike detection with a kohonen feature map. *Annals of Biomedical Engineering*, 28(11):1362–1369, 2000.

[310] Vivek Prakash Nigam and Daniel Graupe. A neural-network-based detection of epilepsy. *Neurological Research*, 26(1):55–60, 2004.

[311] Elif Derya Übeyli. Wavelet/mixture of experts network structure for eeg signals classification. *Expert systems with applications*, 34(3):1954–1962, 2008.

[312] Elif Derya Übeyli. Implementing eigenvector methods/probabilistic neural networks for analysis of eeg signals. *Neural networks*, 21(9):1410–1417, 2008.

[313] Elif Derya Übeyli. Lyapunov exponents/probabilistic neural networks for analysis of eeg signals. *Expert Systems with Applications*, 37(2):985–992, 2010.

[314] Vairavan Srinivasan, Chikkannan Eswaran, and Sriraam. Artificial neural network based epileptic detection using time-domain and frequency-domain features. *Journal of Medical Systems*, 29(6):647–660, 2005.

[315] Marleen C Tjepkema-Cloostermans, Rafael CV de Carvalho, and Michel JAM van Putten. Deep learning for detection of focal epileptiform discharges from scalp eeg recordings. *Clinical neurophysiology*, 129(10):2191–2196, 2018.

[316] Alexander Rosenberg Johansen, Jing Jin, Tomasz Maszczyk, Justin Dauwels, Sydney S Cash, and M Brandon Westover. Epileptiform spike detection via convolutional neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 754–758. IEEE, 2016.

[317] John Thomas, Luca Comoretto, Jing Jin, Justin Dauwels, Sydney S Cash, and M Brandon Westover. Eeg classification via convolutional neural network-based interictal epileptiform event detection. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3148–3151. IEEE, 2018.

[318] Nurettin Acır and Cüneyt Güzeliş. Automatic spike detection in eeg by a two-stage procedure based on support vector machines. *Computers in Biology and Medicine*, 34(7):561–575, 2004.

[319] Nurettin Acir, Ibrahim Oztura, Mehmet Kuntalp, Baris Baklan, and Cuneyt Guzelis. Automatic detection of epileptiform events in eeg by a three-stage procedure based on artificial neural networks. *IEEE Transactions on Biomedical Engineering*, 52(1):30–40, 2005.

[320] Clodoaldo AM Lima, André LV Coelho, and Marcio Eisencraft. Tackling eeg signal classification with least squares support vector machines: A sensitivity analysis study. *Computers in Biology and Medicine*, 40(8):705–714, 2010.

[321] Daniel Kelleher, Andriy Temko, Derek Nash, Brian McNamara, and William Marnane. Svm detection of epileptiform activity in routine eeg. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 6369–6372. IEEE, 2010.

[322] JD Martinez-Vargas, LD Avendano-Valencia, E Giraldo, and G Castellanos-Dominguez. Comparative analysis of time frequency representations for discrimination of epileptic activity in eeg signals. In *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*, pages 148–151. IEEE, 2011.

[323] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.

[324] Nihal Fatma Güler, Elif Derya Übeyli, and Inan Güler. Recurrent neural networks employing lyapunov exponents for eeg signals classification. *Expert systems with applications*, 29(3):506–514, 2005.

[325] Elif Derya Übeyli. Combined neural network model employing wavelet coefficients for eeg signals classification. *Digital Signal Processing*, 19(2):297–308, 2009.

[326] Zafer Iscan, Zümray Dokur, and Tamer Demiralp. Classification of electroencephalogram signals with combined time and frequency features. *Expert Systems with Applications*, 38(8):10499–10505, 2011.

[327] Cheng-Wen Ko and Hsiao-Wen Chung. Automatic spike detection via an artificial neural network using raw eeg data: effects of data preparation and implications in the limitations of online recognition. *Clinical neurophysiology*, 111(3):477–481, 2000.

[328] Miroslaw Latka, Ziemowit Was, Andrzej Kozik, and Bruce J West. Wavelet analysis of epileptic spikes. *Physical Review E*, 67(5):052902, 2003.

[329] Jessica Askamp and Michel JAM van Putten. Mobile eeg in epilepsy. *International journal of psychophysiology*, 91(1):30–35, 2014.

[330] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

[331] Andrew J Gabor and Masud Seyal. Automated interictal eeg spike detection using artificial neural networks. *Electroencephalography and clinical Neurophysiology*, 83(5):271–280, 1992.

[332] Hansjerg Goelz, Richard D Jones, and Philip J Bones. Wavelet analysis of transient biomedical signals and its application to detection of epileptiform activity in the eeg. *Clinical electroencephalography*, 31(4):181–191, 2000.

[333] Giancarlo Calvagno, M Ermani, Roberto Rinaldo, and Flavio Sartoretto. A multiresolution approach to spike detection in eeg. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 6, pages 3582–3585. IEEE, 2000.

[334] M Nuh, Ach Jazidie, and MA Muslim. Automatic detection of epileptic spikes based on wavelet neural network. In *Circuits and Systems, 2002. APCCAS'02. 2002 Asia-Pacific Conference on*, volume 2, pages 483–486. IEEE, 2002.

[335] Piotr J Durka. Adaptive time-frequency parametrization of epileptic spikes. *Physical Review E*, 69(5):051914, 2004.

[336] Malek Adjouadi, Danmary Sanchez, Mercedes Cabrerizo, Melvin Ayala, Prasanna Jayakar, Ilker Yaylali, and Armando Barreto. Interictal spike detection using the walsh transform. *IEEE Transactions on Biomedical Engineering*, 51(5):868–872, 2004.

[337] Guanghua Xu, Jing Wang, Qing Zhang, and Junming Zhu. An automatic eeg spike detection algorithm using morphological filter. In *Automation Science and Engineering, 2006. CASE'06. IEEE International Conference on*, pages 170–175. IEEE, 2006.

[338] Themis P Exarchos, Alexandros T Tzallas, Dimitrios I Fotiadis, Spiros Konitsiotis, and Sotirios Giannopoulos. Eeg transient event detection and classification using association rules. *IEEE Transactions on Information Technology in Biomedicine*, 10(3):451–457, 2006.

[339] AT Tzallas, PS Karvelis, CD Katsis, DI Fotiadis, S Giannopoulos, and S Konitsiotis. A method for classification of transient events in eeg recordings: application to epilepsy diagnosis. *Methods of Information in Medicine*, 45(06):610–621, 2006.

[340] Guanghua Xu, Jing Wang, Qing Zhang, Sicong Zhang, and Junming Zhu. A spike detection method in eeg based on improved morphological filter. *Computers in biology and medicine*, 37(11):1647–1652, 2007.

[341] Z Hilal İnan and Mehmet Kuntalp. A study on fuzzy c-means clustering-based systems in automatic spike detection. *Computers in biology and medicine*, 37(8):1160–1166, 2007.

[342] KP Indiradevi, Elizabeth Elias, PS Sathidevi, S Dinesh Nayak, and K Radhakrishnan. A multi-level wavelet approach for automatic detection of epileptic spikes in the electroencephalogram. *Computers in Biology and Medicine*, 38(7):805–816, 2008.

[343] Marzia De Lucia, Juan Fritschy, Peter Dayan, and David S Holder. A novel method for automated classification of epileptiform activity in the human electroencephalogram-based on independent component analysis. *Medical & biological engineering & computing*, 46(3):263–272, 2008.

[344] Anup Kumar Keshri, Rakesh Kumar Sinha, Rajesh Hatwal, and Barda Nand Das. Epileptic spike recognition in electroencephalogram using deterministic finite automata. *Journal of medical systems*, 33(3):173–179, 2009.

[345] Yakup Kutlu, Mehmet Kuntalp, and Damla Kuntalp. Optimizing the performance of an mlp classifier for the automatic detection of epileptic spikes. *Expert Systems with Applications*, 36(4):7567–7575, 2009.

[346] Ling Guo, Daniel Rivero, Jose A Seoane, and Alejandro Pazos. Classification of eeg signals using relative wavelet energy and artificial neural networks. In *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, pages 177–184. ACM, 2009.

[347] Christine F Boos, V Pereira Maria do Carmo, et al. Automatic detection of paroxysms in eeg signals using morphological descriptors and artificial neural networks. In *Biomedical Engineering, Trends in Electronics, Communications and Software*. InTech, 2011.

[348] Algimantas Juozapavicius, Gytis Bacevicius, Dmitrijus Bugelskis, and Ruta Samaitiene. Eeg analysis–automatic spike detection. *Nonlinear Analysis: Modelling and Control*, 16(4):375–386, 2011.

[349] Mehdi Radmehr and Seyed Mahmoud Anisheh. Eeg spike detection using stationary wavelet transform and time-varying autoregressive model. *International Journal of Computer Applications*, 83(13), 2013.

[350] Vamsidhar Chavakula, Iván Sánchez Fernández, Jurriaan M Peters, Gautam Popli, William Bosl, Sanjay Rakhade, Alexander Rotenberg, and Tobias Loddenkemper. Automated quantification of spikes. *Epilepsy & Behavior*, 26(2):143–152, 2013.

[351] Jing Zhou, Robert J Schalkoff, Brian C Dean, and Jonathan J Halford. A study of morphology-based wavelet features and multiple-wavelet strategy for eeg signal classification: results and selected statistical analysis. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 5998–6002. IEEE, 2013.

[352] Radek Janca, Petr Jezdik, Roman Cmejla, Martin Tomasek, Gregory A Worrell, Matt Stead, Joost Wagenaar, John GR Jefferys, Pavel Krsek, Vladimir Komarek, et al. Detection of interictal epileptiform discharges using signal envelope distribution modelling: application to epileptic and non-epileptic intracranial recordings. *Brain topography*, 28(1):172–183, 2015.

[353] Peter C Horak, Stephen Meisenhelter, Markus E Testorf, Andrew C Connolly, Kathryn A Davis, and Barbara C Jobst. Implementation and evaluation of an interictal spike detector. In *Image Reconstruction from Incomplete Data VIII*, volume 9600, page 96000N. International Society for Optics and Photonics, 2015.

[354] Sahbi Chaibi, Tarek Lajnef, Abdelbacet Ghrob, Mounir Samet, and Abdennaceur Kachouri. A robustness comparison of two algorithms used for eeg spike detection. *The open biomedical engineering journal*, 9:151, 2015.

[355] John Thomas, Jing Jin, Justin Dauwels, Sydney S Cash, and M Brandon Westover. Automated epileptiform spike detection via affinity propagation-based template matching. In *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pages 3057–3060. IEEE, 2017.

[356] Elham Bagheri, Jing Jin, Justin Dauwels, Sydney Cash, and M Brandon Westover. Classifier cascade to aid in detection of epileptiform transients in interictal eeg. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 970–974. IEEE, 2018.