Machine Learning for Healthcare with a Focus on the Early Diagnosis of Epilepsy and Brain Tumor Detection

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Informatik (12)
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Diyuan Lu
aus Shaanxi, China

Frankfurt am Main 2022
(D30)

vom Fachbereich 12 der

Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan (Dean): Prof. Dr. Martin Möller

Gutachter (Supervisor): Prof. Dr. Jochen Triesch

Prof. Dr. Gemma Roig

Datum der Disputation: 2022.03.03

Abstract

Machine learning (ML) techniques have evolved rapidly in recent years and have shown impressive capabilities in feature extraction, pattern recognition, and causal inference. There has been an increasing attention to applying ML to medical applications, such as medical diagnosis, drug discovery, personalized medicine, and numerous other medical problems. ML-based methods have the advantage of processing vast amounts of data. With an ever increasing amount of medical data collection and large, inter-subject variability in the medical data, automated data processing pipelines are very much desirable since it is laborious, expensive, and error-prone to rely solely on human processing. ML methods have the potential to uncover interesting patterns, unravel correlations between complex features, learn patient-specific representations, and make accurate predictions. Motivated by these promising aspects, in this thesis, I present studies where I have implemented deep neural networks for the early diagnosis of epilepsy based on electroencephalography (EEG) data and brain tumor detection based on magnetic resonance spectroscopy (MRS) data.

In the project for early diagnosis of epilepsy, we are dealing with one of the most common neurological disorders, epilepsy, which is characterized by recurrent unprovoked seizures. It can be triggered by a variety of initial brain injuries and manifests itself after a time window which is called the latent period. During this period, a cascade of structural and functional brain alterations takes place leading to an increased seizure susceptibility. The development and extension of brain tissue capable of generating spontaneous seizures is defined as epileptogenesis (EPG) [1]. Detecting the presence of EPG provides a precious opportunity for targeted early medical interventions and, thus, can slow down or even halt the disease progression. In order to study brain signals in this latent window, animal epilepsy models are used to provide valuable data as it is extremely difficult to obtain this data from human patients. The aim of this study is to discover biomarkers of EPG using animal models and then to find the equivalent and counterparts in human patients' data. However, the EEG features for EPG are not well-understood and there is not a sufficiently large amount of annotated data for ML-based algorithms. To approach this problem, firstly, I utilized the timestamp information of the recorded EEG from an animal epilepsy model where epilepsy is induced by an electrical stimulation. The timestamp serves as a form of weak supervision, i.e., before and after the stimulation. Secondly, I implemented a deep residual neural network and trained it with a binary classification task to distinguish the EEG

signals from these two phases. After obtaining a high discriminative ability on the binary classification task, I proposed to divide further the time span after the stimulation for a three-class classification, aiming to detect possible stages of the progression of the latent EPG phase. I have shown that the model can distinguish EEG signals at different stages of EPG with high accuracy and generalization ability. I have also demonstrated that some of the learned features from the network are clinically relevant.

In the task of detecting brain tumors based on MRS data, I first proposed to apply a deep neural network on the MRS data collected from over 400 patients for a binary classification task. To combat the challenge of noisy labeling, I developed a distillation step to filter out relatively "cleanly" labeled samples. A mixing-based data augmentation method was also implemented to expand the size of the training set. All the experiments were designed to be conducted with a leave-patient-out scheme to ensure the generalization ability of the model. Averaged across all leave-patient-out cross-validation sets, the proposed method performed on par with human neuroradiologists, while outperforming other baseline methods. I have demonstrated the distillation effect on the MNIST data set with manually-introduced label noise as well as providing visualization of the input influences on the final classification through a class activation map method. Moreover, I have proposed to aggregate information at the subject level, which could provide more information and insights. This is inspired by the concept of multiple instance learning, where instance-level labels are not required and which is more tolerant to noisy labeling. I have proposed to generate data bags consisting of instances from each patient and also proposed two modules to ensure permutation invariance, i.e., an attention module and a pooling module. I have compared the performance of the network in different cases, i.e., with and without permutation-invariant modules, with and without data augmentation, single-instance-based and multiple-instance-based learning and have shown that neural networks equipped with the proposed attention or pooling modules can outperform human experts.

Acknowledgments

It has been a great adventure of my life during these five years. For the first time going abroad, I was overwhelmed by the excitement of experiencing a new life, endeavoring a new career, and exploring a new land abroad, as well as the anxiety of leaving home and my family for the first time. I am very grateful for this enjoyable and forever memorable journey in my life. I have learned a lot, changed a lot, grew a lot and I could not have done this without my family, friends, and colleagues.

First, I would like to thank my advisor, Prof. Jochen Triesch, for the valuable opportunity to conduct research in his lab. I appreciate the patience when I was stumbling along the way, his support and guidance, or even a simple "well done!", which meant a great deal. Thank you to my second supervisor Gemma Roig and the committee members for the feedback on my thesis and many inspiring discussions.

Many thanks to all the colleagues and friends that I am lucky to have in Frankfurt. In particular, I would like to thank Max for helping me preparing my research plan and the application for the scholarship, which made everything afterwards possible, my office mates Natalie and Lukas, my group members Bruno, Florence, Alex (Lelaise), Charles, Jan, Markus, Danylo, Tristan, FIAS colleagues Raj, Bettina, Bastian, Sigrid, Alex (Achenbach, for the great IT support) and all FIAS administrative members. I owe many thanks to our FIASCO band, Florence, Lukas, Gustavo, and Fabian. It is an amazing group of people who are passionate about music. I truly appreciate all the musical sessions we had and hope we all keep making music in our lives. I will always cherish all the fun conversations, board games, beers, and parties that we had together and I hope we could extend them into the future.

I worked closely with collaborators from the University Hospital Frankfurt from

the CePTER project (including but not limited to the CePTER project). Thank you to Sebastian, Valentin, Nenad, Venassa, Prof. Felix Rosenow, and Prof. Elke Hattingen for inspiring discussions and valuable input to my work.

My family has relentlessly supported me during my PhD. Deepest thank you to my parents Guang Lu and Qing Sun. They raised me to be kind, strong, independent, and loving. There were times that I felt I was trapped in deep doubt of myself, and they made me see things clearer and face the unknown bravely. Many thanks to my grandparents who are traditional Chinese farmers. They taught me to be honest, hard-working, and caring. It is a heartbreaking that during the past three years I lost my father and my grandfather, two of the most important persons in my life. I only wished that they could see what I have become today and face what is to come with me.

Finally, I would like to express my deepest appreciation to my boyfriend, Gerhard, for his loving care and endless support for the past more than six years and many others to come and his parents Hartmut and Waldtraud for their attentive care and warm hospitality every week for the afternoon coffee.

It is not exaggerating to say that every PhD student has his/her painful struggle at the beginning, and I was surely one of them. I am glad that I made it through owing my thanks to all those ones that I mentioned above and the Chinese Scholarship Council (CSC) for the final support.

Contents

List of Figures

Chapter 1

Introduction

Machine learning (ML) methods have witnessed a dramatic booming in the past few years and have demonstrated their great potential in many fields, such as learning games [2, 3, 4], generating high fidelity images [5, 6], style transferring [7], speech recognition and synthesis [8, 9], and natural language processing [10, 11]. They have been further advanced by the increased computational power, the availability of large and specialized data sets, and deeper theoretical understandings of numerous learning algorithms.

In recent years, there has been a flurry of research efforts addressing the application of ML- and deep learning (DL)-based methods in healthcare. There have been impressive studies undertaken for various medical tasks, such as cardiovascular disease classification [12], skin cancer detection [13], lung cancer diagnosis [14], automatic prognosis for diseases [15], COVID-19 treatment and diagnosis [16], and many other medical tasks.

However, amongst challenges that are ubiquitous across those domains, such as the lack of large amounts of annotated data, applying ML methods to healthcare faces unique difficulties. For example, often the medical data is anonymized due to privacy concerns and, of course, that filters out unique information that may

be useful for ML methods. Secondly, the collected data in healthcare is often limited in quantity, corrupted with missing values, sampled irregularities, and highly variable sampling qualities across different recording sites. This hinders the generalization ability of any ML-based methods. Thirdly, given the limited data from each individual, it is especially difficult to personalize predictions of the ML algorithms, which is of great importance in healthcare.

Data in healthcare is highly heterogeneous and commonly covers images, time series data (audio, video, electrocardiography (ECG), electroencephalography (EEG)), text, etc. This thesis focuses on two medical applications: the early diagnosis of epilepsy and brain tumor detection. In the epilepsy project, we trained neural networks to learn based on a large collection of EEG data from a rodent epilepsy model, with the aim to discover potential EEG signatures that are indicative for the ongoing epileptogenic process. Here, EEG signals were recorded continuously from a rodent epilepsy model. In this epilepsy model, rodents were implanted with a depth electrode and epilepsy was induced by electrical stimulation. However, except for the time stamps at which the data was collected, we did not have any other forms of annotation; this poses the major challenge to any supervised learning tasks.

In the tumor detection project, we learnt to classify tumor and non-tumor tissues with magnetic resonance spectroscopy (MRS) data, which is a sequence data reflecting the biochemical composition of the brain tissue. Here, we were confronted with the lack of data, imbalance of classes, and noisy labeling of samples. Meanwhile, there was a heterogeneous distribution regarding the number of samples from each patient, which posed difficulties to learning due to large individual variability. We have presented all the details and results of these two projects in our publications, listed in Chapter 2.

In this chapter, we would like to start by providing a short summary of recent

advances in deep neural network (DNN) design since deep learning is the fundamental tool used in this thesis. We also describe several techniques developed in the course of the studies to overcome various problems during the training of DNNs. Following this, I shall briefly introduce the basic concepts and data acquisition methods used in these two projects. To this end, I shall provide some of the basics of EEG and the EEG signatures of hippocampal rhythms to enable a basic understanding of the data in the task of epilepsy early diagnosis. Furthermore, I shall provide some basics of the MRS data and its implications in the tumor detection task.

1.1 Deep Learning

Deep Learning (DL) is a subfield of machine learning which is concerned with neural networks with multiple layers of artificial "neurons". These networks learn features from a low and primitive level up to a high and an abstract level from the input and make use of these learned features to reach the target output. For example, such a target can be a label in a supervised learning framework, or future values of the input in an autoregressive framework, or the input data in an unsupervised generative model. DL is inspired by considering how information is processed in a real brain for object detection, language learning, and speech recognition. Taking humans as an example, our central nervous system is an amazing master piece in the sense that billions of neurons are connected in a certain way, learning hierarchical representations of the input from primary sensory layers to higher cortical areas in a very efficient manner. Primary sensory neurons are tuned, i.e., they respond strongly to basic primitive features from the input, while neurons further up in the information processing pathway are tuned to more abstract features. Olah *et al.* (2017) elucidated how the network builds up its understanding

of training data over multiple layers in object detection tasks. They illustrated that the early layers respond strongly to a variety of edges. The subsequent layers are mostly active to different textures and these are followed by layers that mostly respond to different patterns which are diverse and the creative combinations of all previous features. The layers following are highly sensitive to various parts of different objects and, finally, the the layers that are close to the output layer show various object shapes that reflect labels [17].

1.1.1 Biological and Artificial Neural Networks

The anatomy of a neuron is shown in Fig. 1-1. The neuron collects input from other neurons through its dendrites which function as antennae. Subsequently, the collected input accumulates in the cell body. When the signal is above a certain threshold, the neuron will send out a signal, i.e., an action potential, through its axon. The neuron sending out the signal is called the pre-synaptic neuron, whilst the ones receiving this signal are called post-synaptic neurons. This signal is not a linear combination of all the inputs but is gated by the axon hillock of the neuron in a non-linear fashion, i.e., a non-linear activation function. The signal completes its transmission at the synapse where the nerve ending of the presynaptic neuron almost touches the post-synaptic dendrite. On average, a neuron in the human brain has several thousand synapses connecting to other neurons [18].

In order to learn effectively, the synaptic connections need to be adjusted in multiple layers simultaneously, according to various forms of neural modulations and neural plasticity. Homeostatic plasticity is an important and well-studied concept that refers to the process of neurons auto-regulating the strength of synaptic connections to prevent neural circuits from being hyper- or hypo-active [19]. One realization of this mechanism is synaptic scaling which ensures that the strength of the synaptic connections is up- or down-regulated proportionally to the changes

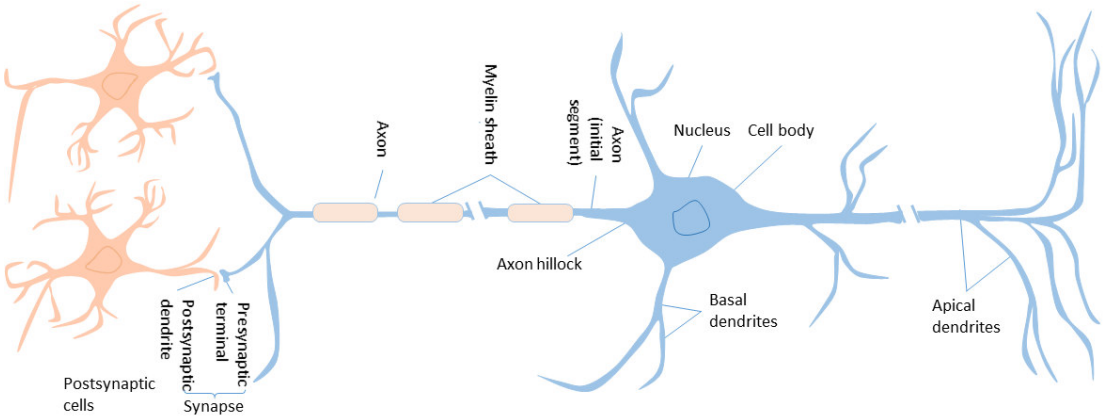


Figure 1-1: Anatomy of a neuron.

in the activity. Thus, this mechanism keeps the network's activity dynamics in a stable regime [19] and is beneficial for learning [20, 21].

Some of the key aspects of biological nervous systems are, to some extent, incorporated into designing "neurons" and "layers of neurons" in artificial neural networks. One of the straightforward network structures is the multi-layer-perceptron (MLP) [22]. This is a class of feedforward neural networks, where multiple layers of neurons are connected sequentially. In an MLP network, the first layer is the input layer, resembling the function of the dendrites of a biological neuron. This receives the input data, be it images, time series data, audio, or text and then passes on the information to the neurons in the next layer. The connections are reflected in the weights from one neuron to another. In each layer, neurons receive inputs from the previous layer and take the sum. Subsequently, they transform the sum with a non-linear activation function, and pass on the signal to the connected neurons in the next layer. Some forms of regularization are beneficial for the network to stabilize its activity, either in a biological neural network or an artificial DNN. For example, the homeostatic plasticity in a biological neural network and the batch normalization, weight normalization, and many of their variants in a DNN [23].

In deep neural networks, the weights of neurons can be adjusted efficiently by an algorithm called backpropagation, in which the error between the output of the network and the target output is propagated back to the very first layer and the weights between neurons are adjusted in such a way that the error is minimized throughout training. However, backpropagation has been viewed as biologically unrealistic; there has been no explicit evidence of such an error being propagated in a biological system. In addition, there have been studies which have attempted to link aspects of the DNNs to those of the biological system. Lillicrap *et al.* (2020) argue that the difference of the activity in feedback connections may provide some form of modulation/supervision when it is approximated by the lower-level circuit [24].

1.1.2 DNN Structures

After building up the correspondence between biological neural networks and artificial neural networks, in the following, we provide a short review on a few popular DNN structures. There have been many successful DNN network structures that have enjoyed great popularity in recent years. One example of the pioneering works is the LeNet neural network, developed by Yann Le Cunin (1998) [25]. This was one of the first convolutional neural networks to be developed and is applied to the modified NIST dataset. This work laid the the foundations for the essential roles that convolution networks now play in numerous applications. It emphasized that image features are often distributed and locally correlated. Convolution can extract image features effectively with locally shared parameters and take the topology of the input into consideration to ensure some extent of shift, distortion, and scale invariance. In 2012, AlexNet was proposed by Alex Kirzhevsky *et al.* [26], which caught much attention after its debut at the ImageNet [27] contest. It extended the idea of LeNet into a larger and deeper structure, which could learn

much more complex objects and represent the features more efficiently. It utilized the Rectified Linear Unit (ReLU) [28] as the activation function in the neural network and achieved faster and better performance compared to networks with *tanh* units. The benefit of ReLU units is that the gradient is constant when the activation is positive. This is especially important in very deep neural networks where the gradient might vanish or explode as the training losses propagate back to the very early layers. In AlexNet, dropout was also applied [29, 30]; this is a type of regularization in the network, to prevent the co-adaptation of feature detectors, where a feature detector is only helpful when working together with several other features detectors, which means it does not detect generally helpful features independently [29].

To capture multi-scale features from images, the Inception network has been proposed and it has gone through several modifications starting from the first version developed by Szegedy *et al.* (2015) [31]. The key design insights are (1) using multi-scale convolutions in parallel in one block to capture features in different scales and (2) applying a 1×1 bottleneck convolution in each convolution branch within the block to reduce feature maps for the future convolutions. This design significantly saves the number of computations needed in large convolutional nets, consequently speeding up the execution time and, thus, achieves efficient feature extraction and recombination. Later on, in Szegedy *et al.* (2016) [32], batch-normalization (BatchNorm) [23] was applied to one variant of the Inception network, referred to as the Inception-v3 model; this model achieved the state-of-the-art performance in the ImageNet classification task. BatchNorm normalizes the feature maps of each layer by subtracting the mean and dividing by the standard-deviation of that feature map such that the response of each feature map is zero-mean and has a variance of one. In this way, the network, without having

any information about the internal covariate offset during training, can only focus on learning the true meaningful structures in the data.

At almost the same time as the development of the Inception-v3 model, another revolutionary network architecture, ResNet, was introduced by He *et al.* (2016) [33]. The idea of the ResNet design is to add "shortcuts" which bypass the input to the output of a multi-layer convolutional block. This makes training networks with a large number of layers possible. Many of the aforementioned network design components are also used in ResNet, such as batch-normalization [23], dropout [30], and ReLUs [28], to achieve state-of-the-art performance. The DenseNet model, an extreme application of "shortcuts", connects every layer directly with each other [34], thus, learned features in each layer can be reused multiple times in other layers and hence decrease the redundancy in the feature maps. A hybrid-inception-ResNet network, Inception-v4, was proposed by Szegedy *et al.* (2017) [35] which exploits the properties from both inception networks as well as residual networks; this also yieldes state-of-the-art performance in the ImageNet classification challenge. ResNet-based networks have demonstrated impressive performances in various tasks including large-scale image recognition [36], medical image segmentation [37], and cardiovascular disease classification [12].

Admittedly, images are ubiquitous and the majority of network structures are designed specifically to learn from images. However, sequential data such as audio, video or text, are also important in real-life applications. Here we review a few successful networks that focus on sequential data. In audio synthesis, WaveNet, proposed by van den Oord *et al.* (2016) [38], learns to generate audio waveforms with a high sampling rate directly from raw audio input in an completely autoregressive fashion. It takes advantage of diluted convolutions in sequence to achieve large receptive fields and generate audio data points that influenced by all previous

data points. Recurrent neural networks (RNNs) have been proposed to deal with sequential data, such as text in [39] and have been successfully applied in numerous tasks including language modeling [40, 41, 42], image captioning [43, 44], speech recognition [45, 46], and machine translation [47]. The main difference between RNN units and convolutional units is that the former receive not only the input from previous layers but also the states from their past time steps, i.e., they have memory of what has been shown and learned. In this way, the network can learn more efficiently with the knowledge unrolled from the past time step, which renders RNN extremely suitable for sequence data including audio, video, and text. The well-known long-short-term-memory (LSTM) units [48, 49] have further equipped the RNN units with several memory gating mechanisms and, as such, the LSTM units are capable of learning long dependencies compared to normal RNN units which often suffer the problem of exploding and vanishing gradients [50].

However, as the length of the sequences, as well as the dependencies, become longer, the RNN-based models become more difficult to train and the performance is found to deteriorate. To address this issue, Bahdanau *et al.* (2014) introduced an attention mechanism which attempts to encode the input sentence into latent vectors, whilst only choosing a subset of these vectors for decoding the translation adaptively [51]. The attention mechanism has been also successfully implemented in various other network structures. The Transformer network, an attention-based model for a sequence-to-sequence mapping task, was first proposed by Vaswani *et al.* (2017) [10] for machine translation tasks. In a sequence-to-sequence mapping task, instead of encoding the information of the input sequence into one hidden state for decoding it in the next step, the attention module computes a mask over all relevant features and hidden states. It then selects the features that are important and ignores background inessential features by element-wise multiplication. The attention mechanism has successfully been applied to vari-

ous tasks including translation [52], building language models [11], and biomedical language mining [53]. Thus, attention models can not only be applied to RNNs, but also to other network structures. In our work, we implemented ResNet-based DNNs, inspired by the work from Hannun *et al.* [12], for EEG classification in Study I [O1], II [O2], and III [O3]. Meanwhile, when doing the MRS classification, we implemented several DNN structures including a fully-connected neural network, a ResNet variant, an Inception net, and a recurrent neural network in Study IV [O4], V [O5], and VI [O6]. Furthermore, in Study VI, we grouped multiple samples from the same patient as a data "bag" taking inspiration from the multiple instance learning framework [54, 55, 56], and an attention module was proposed to handle the permutation invariance within the data "bag".

Furthermore, there is a whole field, called network architecture search, which focuses on searching and designing neural networks automatically when given certain tasks. However, reviewing the current research in this domain is out of the scope of this thesis. For detailed reviews, we refer the reader to [57, 58, 59]. For gaining a general understanding of machine learning and in order to become an expert in this field, we highly recommend the book titled *Machine Learning Yearning* by Andrew Ng [60].

The success of applying DNNs to a real life task largely relies on the understanding of the data at hand. A deeper knowledge of how the data is generated, collected, processed, and the properties of different data modalities is the first stepping stone. In the following sections, we will give a brief overview of the data acquisition methods of both projects.

1.2 Electroencephalography

In this chapter, we provide some basics on electroencephalography (EEG) which is one of the most common procedures to record ongoing large-scale neuronal population activity. Here, we give a general overview of EEG physiology, ongoing hippocampal rhythms, and putative EEG biomarkers for epileptogenesis. Epileptogenesis is defined as the the development and extension of brain tissues capable of generating spontaneous seizures, resulting in either the development of an epileptic condition and/or the progression of the epilepsy after it has been established [61, 1].

EEG has become an important medical data acquisition tool to capture electrical activities of large neuron populations and, thus, can provide insights on normal and abnormal brain signals. It has a very high temporal resolution, in the order of

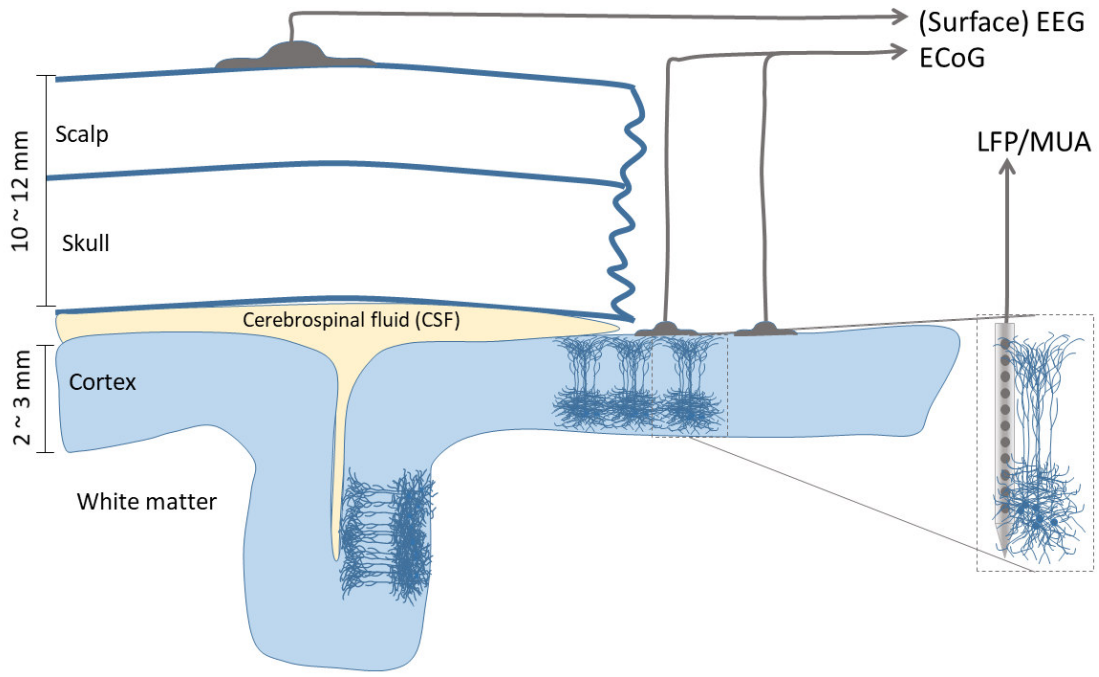


Figure 1-2: Schematic of EEG signal measurements. EEG: electroencephalography; LFP: local field potential; MUA: multi-unit activity; ECoG: Electrocorticography

milliseconds. When neurons are activated, they generate action potentials which, thus, generate electrical currents; these currents produce electrical and magnetic fields. The fields may be recorded by a variety of electrodes that are close to the sources (the local EEG or local field potentials, LFPs), or on the surface of the cortex (the electrocorticogram, ECoG), or on the surface of the scalp (surface EEG) [62]. Figure 1-2 shows different recording methods. In general, electrodes can only measure the activity from a large population of neurons and, depending on where the electrodes are placed, there are more or less distortions of the signals due to the propagation through other tissues.

The brain activity in the hippocampus is of special interest of us, because the hippocampus is an essential component involved in many cognitive functions involving memory formation and consolidation [63, 64, 65] as well as planning and learning [66].

Closely connected to many other cortical areas, the hippocampus is crucial for many essential cognitive functions. Damage to the hippocampus can lead to various neurological disorders, such as Alzheimer's disease, and temporal lobe epilepsy (TLE). Nearly 50% to 75% of epilepsy patients may have hippocampal sclerosis, which is a condition with severe neuronal cell loss in the hippocampus [67]. In the project for early diagnosis of epilepsy, an animal model of mesial-TLE [68, 69] is used and the EEG was recorded from the granule cell layer in the dentate gyrus with a depth electrode, shown as the black star in Fig. 1-3. For simplicity, we use the term EEG to refer to local EEG rather than surface EEG, unless specified otherwise.

Moreover, nearly 50% to 75% of epilepsy patients may have hippocampal sclerosis, which involves severe neuronal cell loss in the hippocampus [67]. Understanding the structure and activity of the hippocampus is one important aspect in our research. There are a few normal hippocampal rhythms that are crucial for various

cognitive functions. When undergoing pathological changes, these rhythms will be altered. Thus, through investigation of the changes of hippocampal rhythms, we could learn more about the ongoing disease.

1.2.1 Hippocampal rhythms

In this section, we give an overview of several well-studied brain rhythms that are present in the hippocampus. The hippocampus is an important component of the brain of humans and other vertebrates, embedded in the deep temporal lobe [67]. It consists of several import regions, i.e., Cornu ammonis-1 (CA1), CA2, CA3 and the dentate gyrus, shown in Fig. 1-3. There are mainly two information pathways in the hippocampus: (1) the perforant pathway: layer-II neurons of entorhinal cortex (EC) $\rightarrow$ the granule cells in the dentate gyrus region $\rightarrow$ the CA3 region $\rightarrow$ the CA1 region $\rightarrow$ back to the EC, (2) the temporoammonic pathway, which is the direct projection from layer-III neurons of EC to the CA1 region of the hippocampus.

Broadly speaking, the neurons in the hippocampus or cortical areas can be divided into two major groups: principal neurons and non-principal neurons or interneurons [70]. Principal neurons are excitatory neurons, which constitute the majority of neurons in the brain. They have long axons, passing information through to other brain areas and activate the down-stream neurons. For example, granule cells in CA1, with long mossy fiber, and pyramidal cells in CA1 and CA3 areas are all principal neurons. Non-principal neurons or interneurons are inhibitory cells. They have dense and local axons, which enables them to moderate, coordinate, and control a large population of neurons locally [70].

It is conventional to discuss brain rhythms in different frequency bands, since the time series data is highly dynamic and often with a low signal to noise ratio. The frequency domain provides a more general and robust overview of the signal.

Furthermore, the generation mechanism of different rhythms in hippocampus is well-studied, thus frequency analysis could provide an overview of the activity from different neuronal assemblies. We approached the data analysis of EEG from the frequency domain, aiming to find frequency features that could be used for the early diagnosis of epilepsy. In this thesis, we mostly focus on the following rhythms as stated in [71]: delta rhythms ($\sim$ 0.5–4 Hz), theta rhythms ($\sim$ 4–12 Hz), beta rhythms ($\sim$ 12–25 Hz), gamma rhythms ($\sim$ 25–100 Hz), sharp-wave ripple complexes ($\sim$ 110–250 Hz ripples superimposed on $\sim$ 0.01–3 Hz sharp waves), and high frequency oscillation ($\sim$ 80–500 Hz). A few waveform examples from different frequencies are shown in Fig. 1-4. Buzsáki *et al.* (2004), provided an overview of

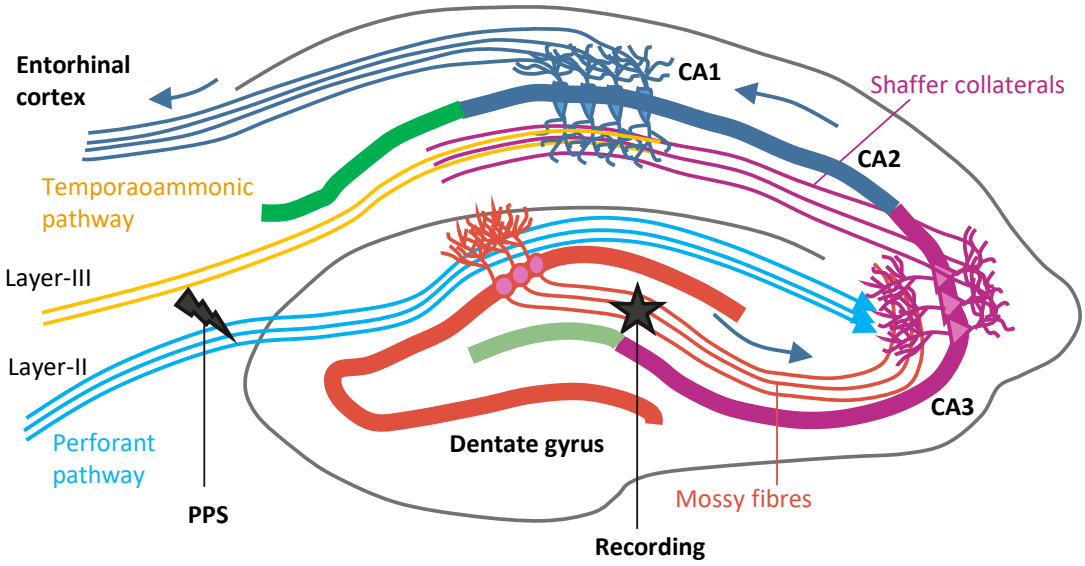


Figure 1-3: A simple schematic of the hippocampal circuitry. There are mainly two pathways in the hippocampal circuit. 1. The axons of layer-II neurons in the entorhinal cortex (EC) $\rightarrow$ the dentate gyrus $\rightarrow$ pyramidal neurons in CA3 through mossy fibers $\rightarrow$ CA1 pyramidal neurons through Schaffer collaterals $\rightarrow$ deep-layer EC neurons. 2. EC layer-III neurons $\rightarrow$ CA1 pyramidal neurons through temporoammonic pathway (TA). CA: Cornu ammonis. The black start indicates the recording site in our experiment. The black lighting symbol indicates the stimulation site. PPS: perforant pathway stimulation.

the frequency bands of brain rhythms, including a range of frequency bands with error bars, concluding that the definition of the frequency bands is not fixed across studies [72]. For more details on brain rhythms, we refer the reader to the book by Buzsáki [66] (2019), where brain rhythms are discussed in great detail within the context of neural functions.

Sharp-wave Ripples (SWRs, $\sim$ 0.01–3 Hz sharp waves superimposed by $\sim$ 110–250 Hz ripples [71]) are deemed to be associated with the information transfer mechanism from the hippocampus to the neocortex when there is no extrinsic input to the network, for example, in non-REM (rapid eye movement) sleep [73, 74]. They reflect the excitation of the CA1, CA3 pyramidal neurons by the synchronous bursting from CA3 pyramidal cells during awake immobility and slow-wave sleep [75, 76]. They have been suggested to be associated with memory consolidation [65, 64], but also in certain aspects during active navigation [63, 77]. Global interference with SWRs could lead to the memory impairment and instability in the spatial representation coding [78].

In physiological hippocampal SWRs, CA1 pyramidal cells fire selectively, i.e., only triggered by specific events or cell-specific drives [76]. However, in pathological SWR generation, pyramidal cells are firing more often and in a nonspecific way, which can be reflected by disorganized spectral features of these SWRs [76].

The Delta Rhythm ($\sim$ 0.5–4 Hz) is often associated with sleep and deep anesthesia, when the network sustains a slow-patterned network activity, even in the absence of sensory input [79, 80]. Delta waves are related to locomotive behavior in rodents [81], where the synchronization of the delta-band develops rapidly during the brief pauses between runs, as well as occurring throughout long stationary bouts. The phase of Delta oscillations also modulates the amplitude of gamma-band activity, which allows the information to be processed in an orga-

nized manner [82, 73]. In the delta oscillation, neurons fluctuate between a period of intense synaptic activity (Up state), and a period of silence (Down state) [80]. The cortical network alone is sufficient to generate and sustain the delta rhythm; especially the cortical layer 5 pyramidal neurons are considered to play a key role in the delta rhythm generation [80]. In the hippocampus, the delta rhythm could come from the direct projections from entorhinal cortex to dentate granule cells, CA1 pyramidal neurons, and interneurons, shown in Fig. 1-3. During the "Up" state, CA1 neurons are activated either by direct input from the entorhinal cortex or by the dentate CA3-CA1 circuit; whereas, during the "Down" state, the self-organized activity in the CA3 region is the main driving force of CA1 neurons [83]. It is also reported that an increase of the delta rhythm occurs during prolonged periods of wakefulness, where groups of neurons go briefly silent as they do during sleep [84]. An example of a delta EEG trace is shown in Fig. 1-4.

Costa *et al.* showed that in epileptic rats, the delta band power increased in animals with high occurrence of generalized seizure. This delta oscillation contributes to the promotion of a large scale recruitment of neurons and ultimately seizures [85]. An increase of delta band power has also been reported in human epilepsy patients [86, 87]. However, the origin of epileptiform delta waves is still unclear.

The Theta Rhythm ($\sim$ 4–12 Hz) is a relatively low frequency sinusoidal signal that is largely involved in spatial navigation and episodic memory [88, 89], cortical modulation, information processing [90], learning and decision making. It is also involved in the representations of remembered, ongoing, or imagined future experiences [91, 90], active exploration and locomotion [92, 93, 88, 82, 94]. In particularly, the 6–8 Hz theta rhythm and high gamma oscillations dominate the dentate region, shown in Fig. 1-3, whereas during waking, there are, predomi-

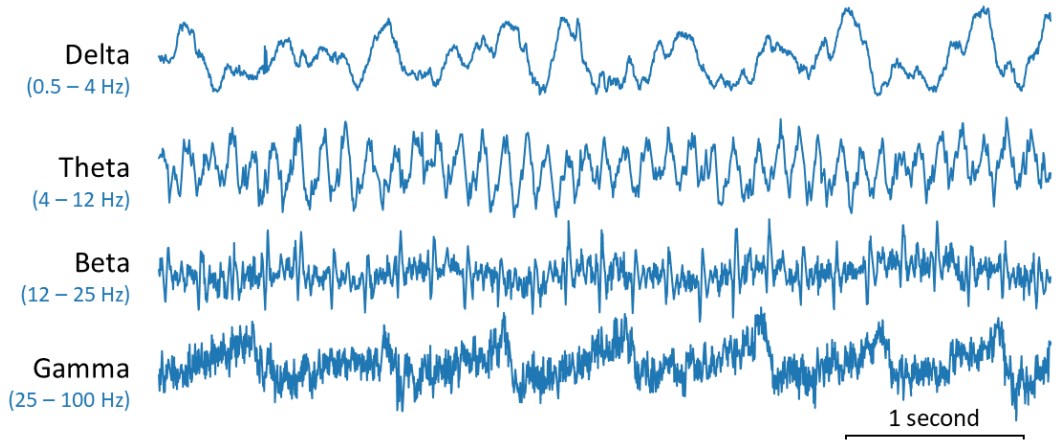


Figure 1-4: Examples of EEG traces at different frequency bands in our rodent epilepsy model. In the last trace, we can see that the gamma rhythm is superimposed on a delta rhythm. Reprinted from our rodent epilepsy model.

nately, 8–10 Hz theta waves and gamma activity in this region [94]. Interneurons in CA1 area also receive rhythmic input from interneurons in the septum, and this septal disinhibition promotes the theta generation in CA1 pyramidal cells [95]. An example EEG trace of the theta rhythm is shown in Fig. 1-4.

It has been accepted for decades that the theta frequency encodes the running speed and can be used to estimate displacement. However, this has recently been challenged by Kropff *et al.* (2021) [96]. By clamping the running speed at a predefined value, they found that the theta frequency is linearly related to positive acceleration and not speed, as previously believed.

Compromised theta activity, however, could reflect cognitive dysfunctions and pathological alterations of the brain. In animals with epilepsy-induced injuries, a decreased tendency to perform exploration in the environment is also observed, which could lead to a decrease of theta activity [97]. For example, in a animal model of temporal lobe epilepsy with pilocarpine injection (an epilepsy-inducing drug), the authors suggested that a decreased ability to generate theta activity may lead to a persistent deficit of spatial memory [97].

The Beta Rhythm ($\sim$ 12–25 Hz) is known to be associated with multiple-modality-input coordination [98, 99]. In the hippocampus, beta oscillations can be generated locally when CA1 principal neurons are being recruited in two alternating gamma periods [100]. Interneurons that receive both feedforward excitation from the CA3 area and feedback excitation from local CA1 neurons can generate beta rhythms in the CA1 area [100]. It is also suggested that beta oscillations in the dentate gyrus are driven by the input from the entorhinal cortex, which modulates functional coupling between hippocampus and other brain areas, thus contributing to the object-position associative learning [101]. Iwasaki *et al.* showed that beta oscillations in the CA1 region were enhanced while the animal was exploring a novel environment, which suggests that these exploration-induced beta oscillations might be a result of multiple cognitive processes such as attention, curiosity, and novelty encoding of the environment [102]. They also found that the beta band power is positively correlated with the performance of mice in a memory-retrieval task, which means the beta rhythm generated during novelty detection also contributes to correct memory acquisition [102]. França *et al.* also reported that the beta band power was enhanced during the exploration of a novel environment and decreased when the animal gets accustomed to the environment setup [103].

The Gamma Rhythm ($\sim$ 25–100 Hz) is often subdivided into "low gamma" ($\sim$ 25–55 Hz) and "high gamma" ($\sim$ 60–100 Hz). Inhibitory interneurons contribute to gamma generation in the hippocampus [104]. To be specific, the "low gamma" rhythm is likely to be driven by CA3-activated interneurons and the "high gamma" is likely to be driven by entorhinal cortex activated interneurons [71]. The "high gamma" has been shown to encode current sensory input and ongoing trajectories [105, 106]. As for the "low gamma", it is hypothesized to be asso-

ciated with memory retrieval through the coupling with the phase of the theta rhythm [107, 108]. The gamma rhythm could be generated by the self-organized CA3-CA1 circuitry independent of cortical-entorhinal inputs and it is modulated by the delta rhythm [83]. The fast-spiking interneurons in the hippocampus can also generate the gamma rhythm providing a gamma-base membrane potential change for other principal cells [100, 109]. Often, the gamma amplitude and gamma phase are coupled to the phase of the theta rhythm [71]. For example, a significantly stronger gamma rhythm is found during theta-associated behaviors, such as rapid-eye-movement sleep, exploring, and navigating than during other non-theta-associated behaviors, such as immobility, grooming, and slow-wave sleep [109]. It is observed that the gamma band power in the CA1 pyramidal layer is phase-locked to the theta band power in this area, and this coherence spans both hemispheres [109].

High Frequency Oscillations (HFOs) include a wide range of frequency components of between 80 and 500 Hz. HFOs between 80–200 Hz can be recorded from the normal hippocampus and entorhinal cortex, but are not present in the dentate gyrus [110]. HFOs reflect the neuronal activity of interneurons when facilitating information transfer over multiple areas [111, 112]. Fast ripples (FRs, 250–600 Hz) are deemed to be pathological and are generated by abnormally bursting neurons. They are often detected both in human patients with mesial temporal lobe epilepsy (TLE) and rodent m-TLE models [112], hence, fast ripples have been proposed as biomarkers for epileptogenesis [110, 112, 113].

Dentate Spikes (DSs) are large-amplitude, short-duration (< 40 ms) activities that are distinct from the background of the hilus of the dentate gyrus, shown in Fig. 1-3. They are associated with the synchronized activity of interneurons and granule cells in the hippocampus [109]. They serve to decrease the network

excitability of the CA3 recurrent circuitry in the intact brain [114] and occur sparsely during behavioural immobility and slow-wave sleep [114]. During DSs, the firing rate of the granule cells in the dentate gyrus is increased while that of pyramidal cells in the CA1 area is decreased [115]. Lensu *et al.* (2019) suggested that DSs, together with sharp-wave ripples, may be crucial for learning. This has been demonstrated by the experiments that the DS-contingent stimulation to the hippocampus improves the performance in a pattern separation task, where associative learning takes place.

The rhythms mentioned above are commonly observed in the hippocampus, some of which are present in an intact brain and some of which occur in a pathological brain. Understanding the origin, propagation, and interaction between different rhythms is of great importance. During the development of epilepsy, i.e., epileptogenesis, structural alterations could be reflected in the physiological signals, and discovering indicative signatures of these changes would provide a precious window for medical intervention before epilepsy becomes fully-established.

1.2.2 Epileptogenesis

Epilepsy can be triggered by a variety of initial brain insults, while the damage accumulates over the course of weeks, months or even years until the onset of the first spontaneous seizure [116]. This period between the initial insult and the onset of the first spontaneous seizure is often referred to as the "latent" period. This period is crucial for the early diagnosis of epilepsy and for initiating medical intervention. It is safe to say that the earlier some biomarkers can be discovered for epileptogenesis in interictal spikes, seizure thresholds, high frequency oscillations, excitability, and behavioral alterations, the more effective treatment and positive outcome may be [117, 118].

Thus, the possibility of using biomarkers, indicative of the presence of epilepto-

genesis, would alert physicians to issue early medical interventions. Furthermore, if biomarkers could be identified for different epileptic processes at different stages of progression, a more personalized and targeted treatment would be possible [119]. However, data from epileptogenesis in human patients is difficult to acquire as the condition only comes to the attention of medical care after seizures have already occurred. Thus, animal models are actively developed to facilitate the understanding and identification of biomarkers of epileptogenesis.

1.2.3 Rodent Model of Temporal Lobe Epilepsy

Due to the difficulty of acquiring EEG during epileptogenesis from human patients, pre-clinical studies to identify potential biomarkers are best conducted using animal models, in which the timing of the potential epileptic insult can be controlled and the course of the epileptogenic process be monitored. The animal models should also be able to demonstrate their potential for translation to humans [119, 120]. There are several well-studied candidate animal models. Traumatic brain injury (TBI) models with weight-dropping and controlled cortical impact [121, 122] may lead to similar forms of brain damage as found in post-traumatic epilepsy (PTE) in humans. Animal models with chemical injections, such as picrotoxin or bicuculline [123], pilocarpine [124], iron [125], and kainic acid [126, 127] can be easily controlled and regulated regarding the severity of epilepsy, duration of seizures, types of seizure, and the duration of the latent epileptogenic phase, while electrical stimulation models that induce status epilepticus may mimic hippocampal sclerosis in humans [68, 128].

The animal model that we have used in our work is a mesial temporal-lobe epilepsy (m-TLE) rodent model proposed by Norwood *et al.* (2011) [68]. In the experimental paradigm, epilepsy is introduced via perforant pathway stimulation (PPS), shown in Fig. 1-3. For a detailed description and introduction to hip-

pocampus circuitry we refer the reader to the following literature [129, 130]. PPS in rodents can evoke excitation or even seizure in the granule cell layer in the dentate gyrus, shown in Fig. 1-3. When the PPS persists for hours, it can lead to neuronal loss and damage in the hippocampus, which may introduce epilepsy. Thus, this model can result in hippocampal sclerosis, which exhibits similar characteristics to those found in human temporal-lobe epilepsy. Therefore, with a latent phase before the first spontaneous seizure, we can have the opportunity to discover biomarkers for identifying the epileptogenesis phase.

Some structural modifications can be observed in a damaged hippocampus, for example the loss of interneurons and granule cells, mossy fiber sprouting, etc. However, the EEG biomarkers of these alterations are not well-known. It is also unclear to what extent the structural changes could tip the hippocampal system to be epileptic and what EEG biomarkers could serve the purpose of identifying different degrees of progression.

1.2.4 Putative EEG biomarkers for Epileptogenesis

There have been several studies which have sought to find EEG biomarkers for epileptogenesis in different animal epilepsy models.

The theta rhythm is widely involved in multiple cognitive functions and the changes in this frequency band have been considered as biomarkers for epileptogenesis. Milikovsky *et al.* (2017) showed that a decreased theta power can not only be a promising diagnostic biomarker for identifying epileptogenesis, but also a prognostic biomarker for post-injury epilepsy (PIE) as well as a pharmacodynamics biomarker for evaluating the efficacy of anti-epileptic drugs [131]. Chauviere *et al.* showed that animals exhibit deficits in hippocampus-dependent memory tasks, which could relate to the the neuronal loss in the hippocampus that leads to a reduced ability to generate theta rhythm [97].

Delta activities can be frequently observed in focal epilepsy patients during EEG inspections [132], and an increase of delta power has been observed at the seizure onset zone [133]. These characteristics of the delta band have been proposed to be a biomarker of the epileptogenic zone and to predict the seizure onset zone [134]. Huppertz *et al.* (2001) investigated the localization ability of the delta activity and interictal epileptiform discharge for the epileptic focus. They found that the delta activity exhibits high accuracy in localizing the epileptic lesion, i.e., the delta rhythm occurs more frequently near the lesion [135]. Naftulin *et al.* inspected the network activity outside the epileptic focus and found a significant increase of the delta band power, which might contribute to the seizure generation and propagation [136].

High frequency oscillations have also been proposed as epileptogenesis biomarkers in several studies. Cello-Oderiz *et al.* (2017) showed that HFOs can better localize the epileptic focus than sharp waves since the former do not propagate from the epileptogenic regions score [137]. Li *et al.* (2018) [138] and Bragin *et al.* (2004) [110] found that the animals which later developed epilepsy exhibited significantly higher probabilities of HFO occurrence, both in the ripple (100-200 Hz) and fast ripple (200-500 Hz) ranges, than those which did not. Meanwhile, they also found that the sooner HFOs appear after the injection, the higher is the occurrence of spontaneous seizures in the chronic phase and the shorter the latent period becomes.

A few other studies focused on the sleep changes, spikes, and non-linear dynamics of EEG signals. Andrade *et al.* (2017) investigated the role of sleep-wake disturbances in epileptogenesis and found that there is a decrease in the dominant frequency and the duration of sleep spindles in a traumatic brain injury epilepsy model with generalized seizures [139]. Sheybani *et al.* (2018) found that in a mouse model of epilepsy via kainate injection, the spatial propagation of a subgroup of

spikes across the brain can be a reliable indicator of epileptogenesis as well as epilepsy in the chronic phase [140]. Rizzi *et al.* (2019) investigated the non-linear dynamics of EEG signals and found a significant negative correlation of the embedding dimension in the recurrence quantification analysis with the severity of the ongoing epileptogenesis, i.e., the more severe the epileptogenesis, the smaller is the embedding dimension [141].

1.2.5 Contributions

The studies mentioned above all focus on one or a few predefined features of interest and preprocess the data correspondingly. Current advancements in the DL field raises the question whether DL can be used to automatically detect the process of epileptogenesis before the first spontaneous seizure. Hence, our research aims to discover EEG biomarkers for identifying epileptogenesis automatically without handcrafted features either in the frequency domain [O1] or in the time domain [O2, O3]. There are structural and functional alterations of the brain during the latent period of epileptogenesis. However, it is still not well-known how these alterations progress, how we can identify different progression stages, and what EEG features could be representative of different epileptogenesis stages. In Study I, II, and III, we investigate the aforementioned aspects and show that it is indeed possible to detect the presence of epileptogenesis with a DNN. Furthermore, the networks reveal EEG features that are indicative of a developing epilepsy. Our contributions can be summarized as follows.

- In Study I [O1], we demonstrate that modern deep learning techniques can be used to successfully detect epileptogenesis in a rodent epilepsy model prior to the first spontaneous seizure with frequency-domain features of EEG recordings. We further show that EEG features pooled in a long time window can better characterize EEG data in different phases.

- In Study II [O2], we show that training with the time series EEG data directly with a similar DNN can further improve the performance. It is also shown that features learned by the DNN for identifying the presence of epileptogenesis are related to the epilepsy-inducing procedure.
- In Study III [O3], we demonstrate that a similar deep learning approach can be applied to successfully stage epileptogenesis by classifying early vs. late epileptogenesis in the same rodent model with high discriminative and generalization abilities. This could potentially open the door for early diagnosis and early medical interventions.

1.3 Brain Tumor Detection with Magnetic Resonance Spectroscopy

A brain tumor is the abnormal growth of the brain tissue, which can be benign or malignant/cancerous. In particular, tumors originating from abnormal growth of glial cells, i.e., gliomas, account for the majority of malignant brain tumors [142]. In clinical practice, the diagnosis accuracy of brain tumor from MRS data is often limited due to the noise in the data, large variability among patients, and the inter-rater bias. ML approaches that are based on DNNs have demonstrated great potential in a range of tasks, sometimes even outperforming human experts [12, 13]. These successes largely depend on fast and powerful computer hardware, better learning algorithms, and large amount of training data. DNNs can extract high-level features and their correlations for the purpose of interest such as classification of tumor and non-tumor tissue in our case. Taking advantage of this learning capacity of DNNs, if may be possible to obtain a better and, most importantly, a faster screening tool to improve clinical practice.

In a previous study by Hattingen *et al.* (2008), they found that certain chemicals in the brain could be a reliable predictors for tumor progression in a cohort of 45 patients [143]. With the advances in ML methods for medical applications, we have collaborated on this project and aim to apply ML-based methods to investigate the diagnostic properties of MRS data as a whole, instead of only focusing on a few metabolites.

The data used in this study was collected with proton MRS (^{1}H-MRS) from the University Hospital Frankfurt. In the following section, we shall introduce some basics of ^{1}H-MRS and its application to brain tumor detection.

1.3.1 ^{1}H-Magnetic Resonance Spectroscopy

MRS detects radio frequency electromagnetic signals produced by nuclei within molecules. It is widely used to obtain *in situ* concentrations of certain chemicals, i.e., metabolites, in brain tissues for tumor detection [144].

The general principle of MRS is that certain atomic nuclei behave as spinning magnetic bars. When exposed to a strong external magnetic field, these nuclei will interact with the magnetic field. In particular, they will absorb energy coming from the external magnetic field and while relaxing back to the resting state, they will resonate at a certain frequency, often at the scale of a million cycles per second. Different chemicals resonate at different frequencies and these signals can only be distinguished from each other by a few cycles per second. To clearly separate these signals, it is common practice to specify the frequency of a particular chemical and describe others by stating how much its frequency is shifted from that of the standard reference compound (tetramethylsilane for ^{1}H-MRS). In this way, the MRS spectrum is presented by the chemical shift in parts per million (ppm), shown in Fig. 1-5 [145].

Furthermore, the characteristics of the metabolism of different brain tissue

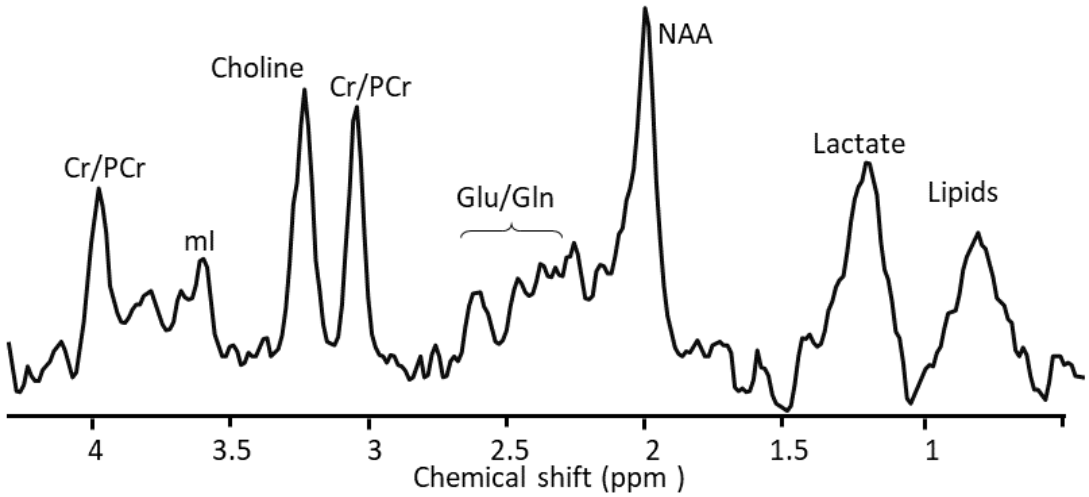


Figure 1-5: An example ^{1}H-MRS sequence with commonly known metabolites. NAA: *N*-acetyl aspartate. Cho: choline; Cr: creatine; PCr: phosphocreatine; mI: myo-inositol; Glu-Gln: glutamate and glutamine compounds.

types vary depending on their functions in the brain. Energy consumption, membrane synthesis and breakdown, cell proliferation, cell loss, etc, all have their footprints in the metabolism profile, especially reflected in some major metabolites such as creatine, choline, lactate, glutamate, glutamine, and *N*-acetyl aspartate (NAA). Understanding how the metabolism profiles differ between different brain tissues is critical in elucidating the underlying pathological conditions [146].

1.3.2 Important Metabolites

As mentioned before, a brain tumor is the abnormal and aggressive growth of brain tissue. It often exhibits excessive energy consumption, rapid cell proliferation, increased cell death, the breakdown of membranes, etc. These characteristics can, thus, be reflected in the profile of the metabolite concentrations in the tissue. Those metabolites that can be detected by a standard ^{1}H-MRS are *N*-acetyl aspartate (NAA), choline, creatine (Cr), myo-inositol (mI), glutamate and glutamine

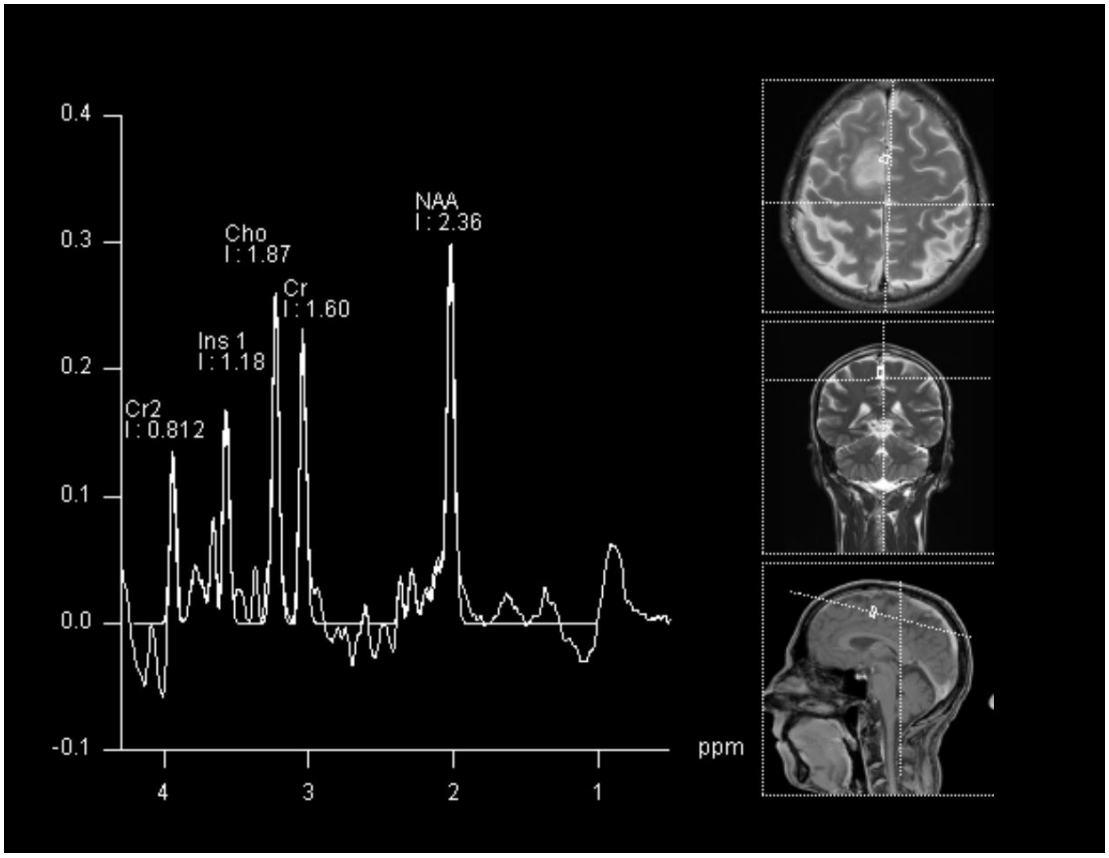


Figure 1-6: An example spectrum from a healthy tissue.

compounds (Glu-n), lipids, and lactate. Figure 1-6 shows an MRS spectrum from a healthy example while Fig. 1-7 shows an example from tumor tissue.

The peak of mI occurs at 3.56 ppm and it is absent from neurons, since it is synthesized in glial cells and cannot pass through the blood-brain barrier [147]. It has been proposed that an increase of the mI level can be indicative of an increase of glial cell size or glial proliferation, both of which can be present in inflammatory processes and some other cerebral diseases [147]. Hattingen *et al.* investigated the role of mI in various glial tumors and reported an elevated level of mI in all glial tumor samples compare to that in non-tumor samples from the same patient [148]. However, due to a heterogeneous distribution of cell composition, cell density, and

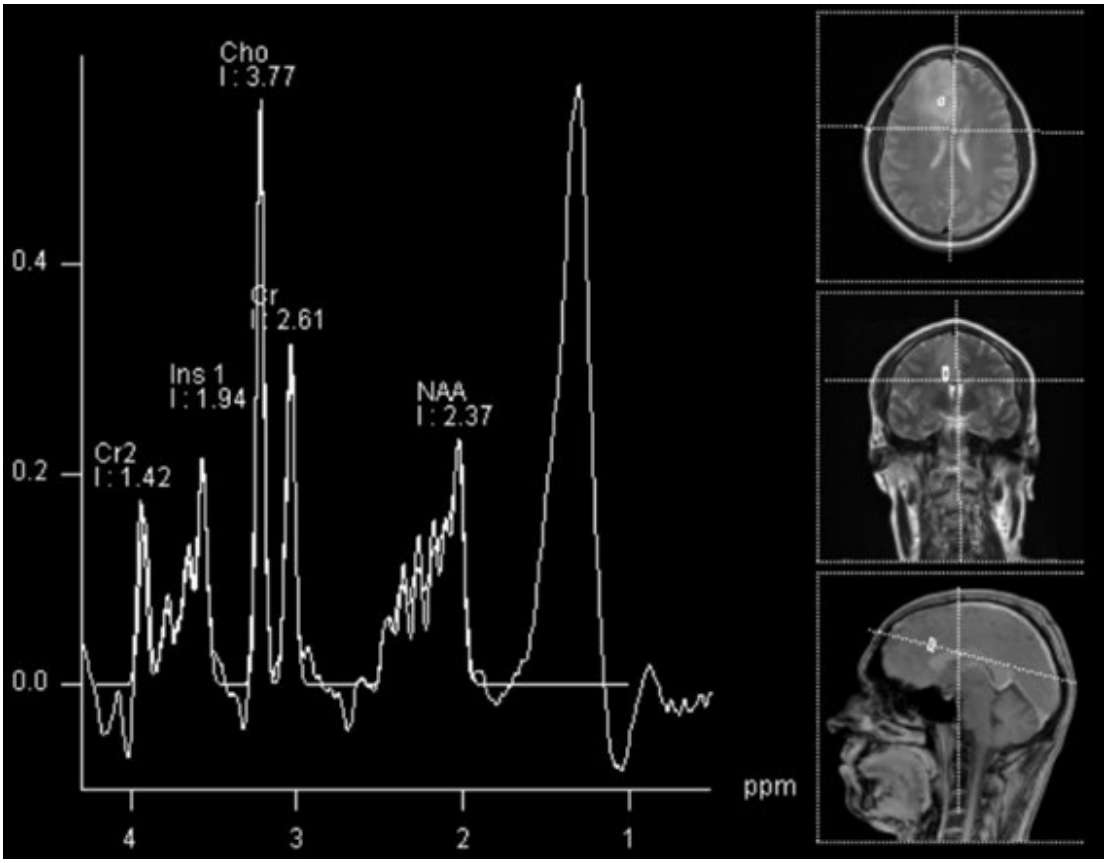


Figure 1-7: An example spectrum from a tumor tissue.

cell proliferation rate, the increase of mI concentration cannot be clearly attributed to certain histopathological processes [148].

The choline peak is located at 3.22 ppm and it reflects the metabolism of the cellular membrane turnover. Hence, it is increased in all processes leading to hyper-cellularity, including accelerated membrane synthesis, which is found in brain tumors [147, 149, 150].

The dual peaks of Cr locate at 3.03 ppm and 3.93 ppm resonant frequencies. It reflects the underlying energy-dependent processes in the brain cells. Cr is involved in energy metabolism, diffusing from the energy producing sites (i.e., the mitochondria) to energy consumption sites (i.e., the nerve terminals) [151, 150].

Cr is not naturally synthesized in the brain, thus, its concentration should be stable across different age groups or even various disease conditions. Therefore, it is convenient to compute the concentration ratios with other metabolites, such as NAA/Cr and choline/Cr [152, 153, 154]. Decreased levels of Cr have been observed in brain tumors [155].

NAA is a marker for neuronal density, which will be reduced in all diseases involving neuron loss or the replacement of neurons by other cells [147]. In healthy brain tissues, the NAA peak is the most prominent peak and locates at 2.0 ppm. It exclusively resides in the central and peripheral nervous systems and reflects the neuronal density and viability [150, 147]. Interestingly, the NAA concentration increases as the brain matures and the Cho concentration decreases [147]. The percentage change in the Cho:NAA ratio has been proposed to be a marker for predicting tumor progression in young brain tumor patients [153]. The NAA:Cr ratio is shown to be helpful in differentiating low- and high-grade gliomas [152].

Glutamate and glutamine (Glu-Gln) together make up a complex of peaks between 2.15–2.5 ppm. In lower-resolution scanners they are difficult to distinguish; only at 3T or higher do they begin to be resolved. Glutamate is the major excitatory neurotransmitter; it is released by neurons during normal brain functions and is then taken up to synthesize glutamine. The glutamine is then transported to neurons to finish the glutamine-glutamate cycle [156]. The Glu-Gln peaks are more detectable and prominent in tumors than in healthy tissues since tumor cells disrupt the uptake of glutamate and strive for more glutamate and glutamine as energy sources that benefit their growth and invasion [156, 157].

The peaks of lactate and lipids resonate at 1.32 ppm and 0.9–1.3 ppm, respectively. In the normal brain, they should be at the threshold of detectability by ^{1}H-MRS. Thus, any detectable increase in lactate and lipids can be viewed as abnormal. Lactate provides an index of metabolic rate and clearance, thus, increased

lactate levels can be observed in conditions where oxygen supply is restricted such as in ischemia [158] and tumors [147, 159, 150, 155].

Lipids in the MRS spectrum result from mobile fatty acyl moieties that no longer bind to the cell membrane during membrane breakdown. An elevated lipid peak may suggest the presence of cerebral tissue destruction, such as in neuronal death [150].

Detecting representative spectrum features for brain tumor and healthy brain tissues forms the focus of this study.

1.3.3 Contributions

Research interest has often focused on several major metabolites, such as NAA, Cr, lipids, lactate, Cho, and some of their ratios, such as NAA:Cr and Cho:Cr ratios [152, 153, 154, 148]. However, tumor diagnosis based on a full-scale landscape of the MRS spectra is lacking. Training DNNs with MRS spectra directly provides the opportunity to extract features across the whole range and capture more complex correlations between different metabolites that normally are not obvious. Furthermore, the diversity in factors such as the tissue composition, cell composition, and cell density leads to diverse appearances of the MRS spectra even for the same class. This imposes immense difficulty in the binary classification task with a single MRS spectrum. We need methods that are more tolerant to noisy labels and form decisions based on the full set of an individual's samples. It is still not obvious whether a DNN-based model trained on the whole range of MRS data could provide a fast and reasonably good screening performance assisting clinical practice. In Study IV, V and VI below, we investigate this possibility and show that indeed our proposed methods could perform on par with or even better than the human experts. Moreover, the networks have learned the traditional well-studied features, mostly concerning individual metabolites or ratios between

two metabolites, as well as discovered new features that prompt further research. Our contributions can be summarized as follows.

- In Study IV [O4], we propose a broad DNN-based framework for classification with noisy labels and scarce data, which consists of a noisy data distillation step and a data augmentation step. When applied to brain MRS data for tumor detection, our model performs on par with neuroradiologists.

- In Study V [O5], following the line of research from Study IV, we show an comprehensive exploration of different network settings and hyperparameters, and provide a rationale for the parameter selection for this task. Towards an explainable AI, we visualize the decision making of the DNN through CAM [160], which shows that conventionally concerned metabolites did show high importance weights.

- In Study VI [O6], we obtain further improvement when applying a multiple instance learning (MIL)-based [54, 55, 56] approach to combat the challenges of noisy labels and data scarcity. An attention-module and a pool-based module are proposed to enforce the permutation invariance in the MIL pipeline. We also provide visualizations explaining the network's learning and decision making process. The proposed method obtains an above-human-expert level performance.

Chapter 2

Publications

This thesis is based on the following papers, which are referred to in the following text by their Roman numerals:

I Deep Residual Neural Network Based Framework for Epileptogenesis Detection in a Rodent Model with Single-Channel EEG Recordings [O1]

II Towards Early Diagnosis of Epilepsy from EEG Data [O2]

III Staging Epileptogenesis with Deep Neural Networks [O3]

IV Human-Expert-Level Brain Tumor Detection Using Deep Learning with Data Distillation and Augmentation [O4]

V Human-Expert-Level Brain Tumor Detection Using Deep Learning with Data Distillation and Augmentation [O5]

VI Multiple Instance Learning for Brain Tumor Detection from Magnetic Resonance Spectroscopy Data [O6]

Own Publications

[O1] D. Lu, S. Bauer, V. Neubert, L. S. Costard, F. Rosenow, and J. Triesch, "A deep residual neural network based framework for epileptogenesis detection in a rodent model with single-channel eeg recordings," in *©2019 IEEE. Reprinted, with permission from D. Lu, S. Bauer, V. Neubert, L. S. Costard, F. Rosenow, and J. Triesch, "A deep residual neural network based framework for epileptogenesis detection in a rodent model with single-channel EEG recordings". 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI).*, pp. 1–6, IEEE, 2019. doi:10.1109/CISP-BMEI48845.2019.8965693.

[O2] D. Lu, S. Bauer, V. Neubert, L. S. Costard, F. Rosenow, and J. Triesch, "Towards early diagnosis of epilepsy from eeg data," in *Machine Learning for Healthcare Conference*, pp. 80–96, PMLR, 2020.

[O3] D. Lu, S. Bauer, V. Neubert, L. S. Costard, F. Rosenow, and J. Triesch, "Staging epileptogenesis with deep neural networks," in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–10, 2020.

[O4] D. Lu, N. Polomac, I. Gacheva, E. Hattingen, and J. Triesch, "Human-expert-level brain tumor detection using deep learning with data distillation and augmentation," in *©2021 IEEE. Reprinted, with permission from*

D. Lu, N. Polomac, I. Gacheva, E. Hattingen, and J. Triesch, "Human-expert-level brain tumor detection using deep learning with data distillation and augmentation". 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3975–3979, IEEE, 2021. doi:10.1109/ICASSP39728.2021.9415067.

[O5] D. Lu, N. Polomac, I. Gacheva, E. Hattingen, and J. Triesch, "Human-expert-level brain tumor detection using deep learning with data distillation and augmentation (submitted)," 2021.

[O6] D. Lu, G. Kurz, N. Polomac, I. Gacheva, E. Hattingen, and J. Triesch, "Multiple instance learning for brain tumor detection from magnetic resonance spectroscopy data," *arXiv preprint arXiv:2112.08845*, 2021.

A Deep Residual Neural Network Based Framework for Epileptogenesis Detection in a Rodent Model with Single-Channel EEG Recordings

Diyuan Lu*, Sebastian Bauer†, Valentin Neubert‡§, Lara Sophie Costard‡ ¶, Felix Rosenow† and Jochen Triesch*
*Frankfurt Institute for Advanced Studies, Frankfurt am Main, 60438, Germany
†Neurology and Epilepsy Center Frankfurt Rhine-Main, University Hospital Goethe-University, Frankfurt am Main, Germany
§Universitätsmedizin Rostock, Oscar-Langendorff-Institut für Physiologie, Rostock, Germany
‡Philipps University Marburg, Translational Epileptology, Marburg, Germany
¶Tissue Engineering Research Group, Royal College of Surgeons Ireland, Dublin, D02, Ireland

***Abstract*—Epilepsy is one of the most common neurological disorders affecting patients across all ages. During the progression of the disease, termed epileptogenesis (EPG), patients may not yet show any clinical manifestation. The EPG phase can range from weeks to years and patients with epilepsy are usually diagnosed by the occurrence of a spontaneous seizure followed by electroencephalography (EEG) monitoring in the hospital. However, the more seizures they have, the less effective the treatment will be. Detecting the development of epilepsy before the first spontaneous seizure may allow for earlier intervention and better treatment outcome. Here we propose a framework based on deep residual neural networks to identify the EPG phase based on EEG recordings in a rodent model where the epilepsy is induced by perforant pathway stimulation (PPS). A deep convolutional neural network is trained to distinguish EEG data recorded before (baseline period, BL) and after (epileptogenesis period, EPG) the EPG is triggered. The proposed model takes the Fast Fourier Transform (FFT) of the preprocessed five-second long EEG segments as input. During testing, we apply a prediction aggregation across multiple consecutive segments to accumulate information over a longer time period. When classifying a continuous stretch of one hour of data, our model achieves $83\,\%$ sensitivity and $83\,\%$ specificity. Further analysis suggests interpretable features in the FFT transformed data that contribute to the distinction of the two phases.**

I. Introduction

Epilepsy is the fourth most common neurological disorder. It is usually accompanied by recurrent seizures and affects more than 65 million people of all ages worldwide [1]. Various forms of acute brain injury can lead to epilepsy and the gradual process of underlying brain structural and functional changes is termed epileptogenesis [2]. Kwan et al. showed that the number of seizure episodes prior to the clinical visit is negatively correlated with the effectiveness of the subsequent treatment [3], suggesting an advantage of earlier intervention. Currently there is no established method to know when the brain first becomes epileptic. Nevertheless, interventions to modulate or even prevent EPG will likely be most successful if applied early during the process [2]. Clinically, EEG is a commonly used tool in epilepsy diagnosis. Hence, discovering EEG-biomarkers to identify patients at high risk of developing epilepsy could be of great value. Currently, inter-ictal epileptiform discharges (IEDs) such as spikes and sharp waves are the only established EEG patterns to suggest an ongoing EPG from human scalp recordings. However, it is unknown when IEDs appear during the course of EPG, and identification of IEDs by visual inspection is subjective. Therefore, developing a system that could reliably detect EPG based on EEG signals would be very useful.

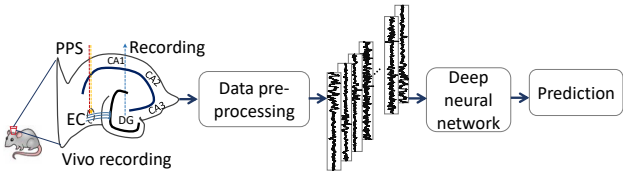


Fig. 1. Proposed framework for epileptogenesis prediction. . EC: entorhinal cortex, DG: dentate gyrus, CA: cornu ammonis, PPS: perforant pathway stimulation.

There have been studies on classifying EEG signals automatically [4], [5]. Conventionally, it is done by extracting hand-crafted features. The method heavily depends on domain expertise. However, in a problem such as EPG identification, little prior knowledge is available on what features to look for [2]. Pre-defining features may result in discarding valuable information that has never made itself obvious.

A. Proposed Framework

In recent years, deep learning techniques have brought revolutionary advances in numerous scientific fields such as speech synthesis, real-time object detection, semantic image synthesis, pedestrian re-identification, natural language processing, as well as many health-care related applications such as heart disease diagnosis, skin disease diagnostic classification and breast cancer detection and classification. One advantage of deep learning based methods is that they learn informative features directly from the data without any human bias. In many domains, such machine-learning-based systems are outperforming humans on difficult tasks [6], [7]. Our goal is to leverage the feature extraction ability of a deep neural

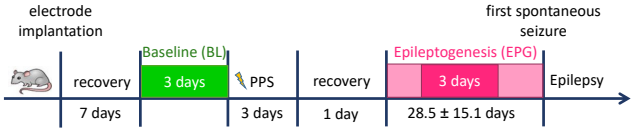


Fig. 2. Timeline of the experiments.

network to distinguish EEG signals recorded in the BL and the EPG periods. The network takes the Fourier transformed preprocessed EEG segments as input. Previous studies have shown that during the EPG phase, the power of certain frequency bands could be enhanced or diminished [8]–[11]. Thus, there might be relevant information in the frequency spectrum to distinguish the two phases. The output of the model is a class label indicating in which phase the segments were recorded. Each segment is rather short (5 seconds) and it may or may not contain relevant information to distinguish the two phases. To pool information across a longer period of time, we propose to aggregate the predictions of multiple segments to obtain the final classification decision. Specifically, we make the following contributions:

- We propose a deep neural network framework to classify FFT transformed EEG data in the context of identifying EPG, which may facilitate the early diagnosis of epilepsy.
- We propose to combine a prediction aggregation of the output of the neural network to pool information across longer time periods.
- We test our approach with EEG data recorded from a rodent epilepsy model with a single-channel depth electrode. The method can distinguish the two phases with a significantly above chance level of performance. The results suggest that there exists considerable cross-subject variability.

II. Material and Methods

Figure 1 illustrates the workflow of our proposed framework. The network takes short segments of EEG signals and outputs a predicted probability over the two possible classes, i.e., BL ("0") and EPG ("1").

A. Data collection and preprocessing

The data we used in this study are collected from a mesial temporal lobe epilepsy with hippocampal sclerosis (mTLE-HS) rodent model, where the epilepsy is induced through perforant pathway stimulation (PPS). We adopted the stimulation paradigm described in detail in [12]. Through the continuous recording in the rodent model we are able to monitor the whole progression of epilepsy and potentially open the door to discover early biomarkers of EPG. We included seven rats undergoing PPS in our analysis. The EPG phase starts with the PPS and ends with the first spontaneous seizure. On average the EPG phase lasted 4 weeks (range 1-7 weeks). The timeline for the PPS-treated rats is shown in Fig. 2. To trade off the computation cost and accuracy, we take three days of the EEG recordings in the BL phase and assign to them the label "0". Likewise, we take three days of recordings in EPG phase and assign the label "1". In total, this gives us more than 980 hours of recordings.

B. Preprocessing

In our setup, the EEG signals are recorded through wireless transmitters with a sampling rate of 512 Hz. A band-pass filter between 0.5 - 160 Hz and a notch filter at 50 Hz were applied to the raw data. Due to electronic interference and unexpected weak transmission, the recordings are partially corrupted and there is some signal loss. To deal with these problems, we first apply an outlier filtering method to filter out unrealistic extreme values. Then, we take non-overlapping five-second segments and applied a data loss filtering where we excluded the ones with over 20% of data loss. At the end we have more than 740,000 segments in total for the experiment.

In this work, we aim to see whether the EEG recordings can be distinguished using frequency information. Hence, we propose to train the network with FFT transformed time series data. The FFT is computed on each five-second segment sampled at 512 Hz. For training, we only take the real part of the FFT, which yields input vectors with 1281 dimensions.

We implement a deep residual neural network with 33 convolutional layers and skip connections [13]. It is inspired by the network architecture in [6]. The concept of residual connections was first proposed by He et al. for an image recognition task [13]. In a deep residual neural network, there are usually multiple residual blocks. One block usually consists of multiple computational layers such as convolutional or dense layers with necessary batch normalization [14], drop-out [15], and a non-linear activation transformation [16]. The input to the residual block is split into two branches: the main branch with all the computations (convolution or dense matrix multiplication, batch-normalization, drop-out) but before the non-linear transformation and another branch usually with identity transformation or max-pooling. The outputs of these two branches are added together and then passed through a non-linear activation function as the input of the next block. In our implementation, there are 15 residual blocks following the classic structure [13]. Each residual block consists of two convolutional layers with necessary batch normalization, drop out and ReLU non-linear activation function in between. The convolutional layers have a filter width of 3. The number of filters increases by a factor of 2 in every four blocks starting from 16. The feature maps were down-sampled in every other block with a stride of 2. To keep the dimensionality compatible, the max-pooling branches share the same stride value as in the second convolutional layer in each block. We apply a dropout rate of 0.2 in all blocks. A soft-max output layer is following the last residual block. Empirical trials showed that with a fully connected layer, the network is more prone to over-fit, which resulted in worse generalization ability to unseen data. So in our model, we leave out the fully-connected layer. The soft-max layer takes the flattened feature maps as the input directly and outputs a probability distribution

TABLE I
NETWORK STRUCTURE USED IN THIS WORK. 3×1 IS THE FILTER SHAPE. 16×2^i IS THE NUMBER OF FILTERS IN EACH BLOCK.

Name	Configuration	Stride	Factor i	Output size
Conv	$\begin{bmatrix} 3\times1, & 16\times2^i \end{bmatrix}$	1	0	[1281, 1, 16]
ResBlock 0	$\begin{bmatrix} 3\times1, & 16\times2^i \\ 3\times1, & 16\times2^i \end{bmatrix}$	1	0	[1281, 1, 16]
ResBlock 1	$\begin{bmatrix} 3\times1, & 16\times2^i \\ 3\times1, & 16\times2^i \end{bmatrix}$	1	0	[1281, 1, 16]
ResBlock 2	$\begin{bmatrix} 3\times1, & 16\times2^i \\ 3\times1, & 16\times2^i \end{bmatrix}$	2	0	[641, 1, 16]
ResBlock 3	$\begin{bmatrix} 3\times1, & 16\times2^i \\ 3\times1, & 16\times2^i \end{bmatrix}$	1	0	[641, 1, 16]
ResBlock 4	$\begin{bmatrix} 3\times1, & 16\times2^i \\ 3\times1, & 16\times2^i \end{bmatrix}$	2	1	[321, 1, 32]
ResBlock (5,..., 8)	$\begin{bmatrix} 3\times1, & 16\times2^i \\ 3\times1, & 16\times2^i \end{bmatrix}$	(1, 2, 1, 2)	(1, 1, 1, 2)	[81, 1, 64]
ResBlock (9,..., 12)	$\begin{bmatrix} 3\times1, & 16\times2^i \\ 3\times1, & 16\times2^i \end{bmatrix}$	(1, 2, 1, 2)	(2, 2, 2, 3)	[21, 1, 128]
ResBlock (13, 14, 15)	$\begin{bmatrix} 3\times1, & 16\times2^i \\ 3\times1, & 16\times2^i \end{bmatrix}$	(1, 2, 1)	(3, 3, 3)	[11, 1, 128]
Dense	2			[2]

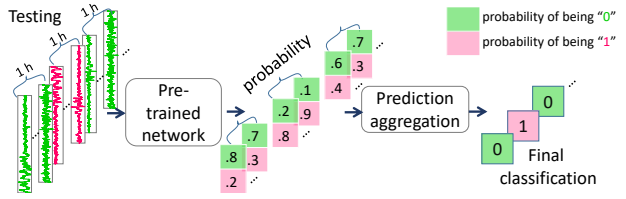


Fig. 3. Test process in our proposed framework with prediction aggregation. See text for details.

over possible classes. The detailed parameters of the network structure are shown in Table I.

III. EXPERIMENTS AND RESULTS

A. Training procedure

To test the generalization ability of our approach, we applied a leave-one-out (LOO) cross validation scheme. We train the network seven times, always leaving out data from one rat which we then use for testing. During training and validation, we uniformly and randomly select 25 hours from each phase from each rat. We adopted a train-validation-split of 9:1, which is a common practice when the size of the dataset is large enough ($\geq 100,000$). The choice of 25 hours is a good trade-off between computation cost and performance from our experience. After the network is trained, we test it with previously withheld three days of data.

B. Prediction aggregation

We propose to aggregate the network predictions of the soft-max output before the categorical predicted label is obtained (Fig. 3). Our method is inspired by the idea of averaging model predictions in a multi-model supervised learning scheme [17], [18]. The intuition is that the magnitude of certainty is taken into consideration. In contrast, a majority vote method would predict the label which has the most counts among the predicted labels. This has the problem of not taking into account the magnitude of certainty of individual clasifications. A sample with 60% certainty contributes equally as a sample with 99.9%.

In our supervised learning scheme the EEG segments from one hour h are $x_{(h,i)}$ and the associated labels are $y_{(h,i)}$, where $i = 1, \ldots, N$ and N is the total number of the samples in this hour. The soft-max output of the model is given by $\hat{y}_{(h,i)} = f(x_{(h,i)}|\text{model})$ and it is in the shape of $[N, 2]$ where 2 is the number of classes in our supervised task setting. The aggregated prediction for hour h is given by:

$$\hat{y}_h = \sum_i^N \hat{y}_{(h,i)} = \sum_i^N f(x_{(h,i)}|\text{model}) \tag{1}$$

and in shape $[1, 2]$. At the last step, we normalize $\hat{y}_h$ along the column axis such that two values sum up to one and can be interpreted as a probability distribution over the two classes after aggregation.

C. Experiments

In the experiment, we perform the classification between signals recorded before and after the PPS, i.e., from BL and EPG phases. Our aim is to see whether the recordings from these two phases can be separated and how separable they are. Clinically, the visual inspection of EEG by experts is often not successful in the EPG phase identification since IEDs appear in both BL (result from the electrode implantation) and EPG phases.

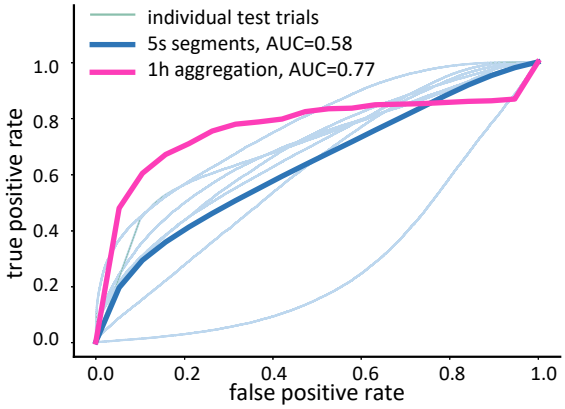


Fig. 4. ROC curves. Individual ROC curves from each LOO test trial (thin light blue). The average ROC curve without aggregation (thick blue) and the average ROC curve with one hour of aggregation (thick pink). ROC: Receiver operating characteristic, AUC: area under the curve.

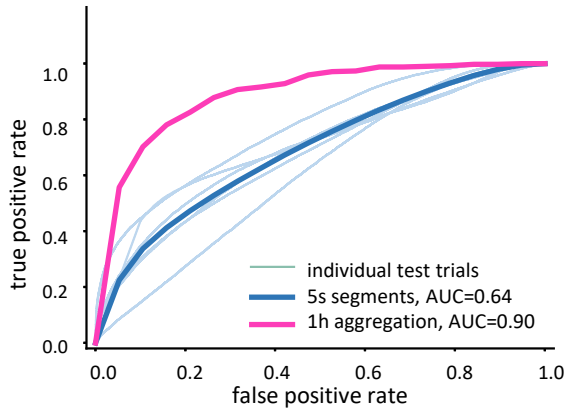


Fig. 5. ROC curves without outlier rat #4. Individual ROC curves from each LOO test trial (thin light blue). The average ROC curve without aggregation (thick blue) and the average ROC curve with one hour of aggregation (thick pink).

D. Results

To simplify the notation, the number of true positives, true negatives, false positives and false negatives are denoted as TP, TN, FP and FN, respectively. The sensitivity, specificity and the area under the curve (AUC) are used to evaluate the classification results.

$$\text{Sensitivity (SEN)} = \frac{TP}{TP + FN}$$

$$\text{Specificity (SPE)} = \frac{TN}{TN + FP}$$

In Fig. 4, we show the results from the Receiver Operating Characteristic (ROC) analysis. We computed the area under the curve (AUC) values in two scenarios: 1) without prediction aggregation, i.e., all five-second segments are independent and contribute equally to the final result. 2) with one hour aggregation where the predictions of all the segments in one hour are pooled together through our proposed method.

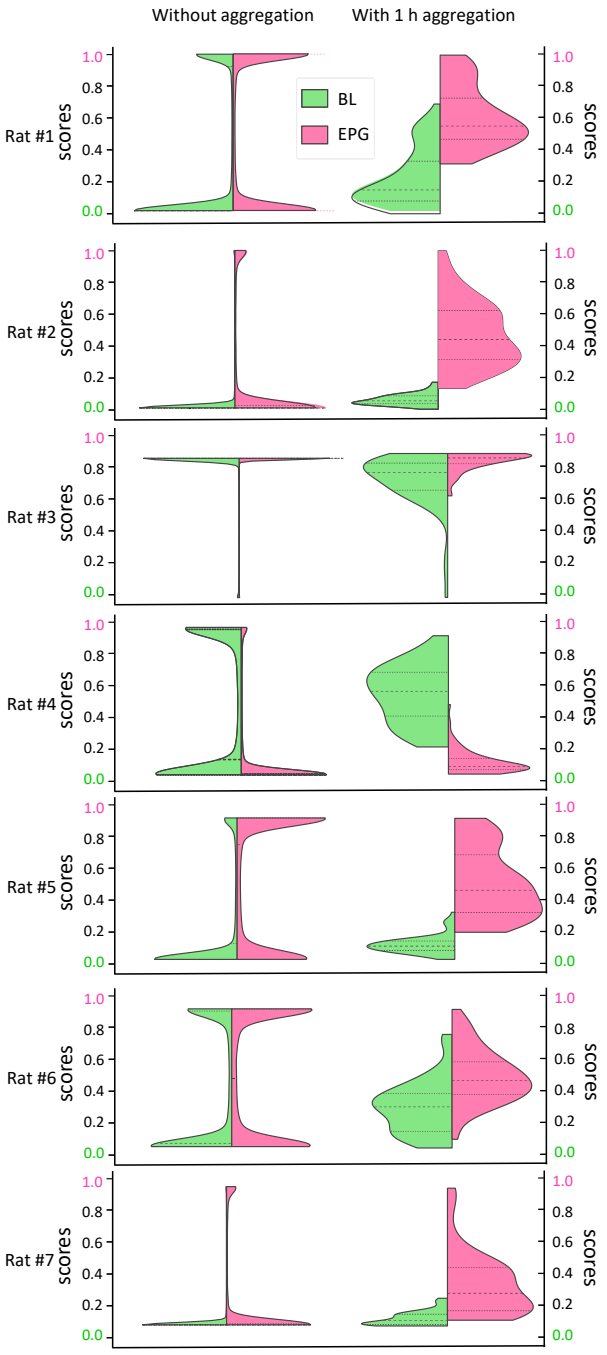


Fig. 6. Distributions of scores of both classes from all test trials. The first column is the distribution without aggregation and the second column is with one hour of aggregation.

The thin light blue lines are from all test trials. Thick lines are averaged ROC curves of all LOO test trials with (thick pink) and without (thick blue) aggregation. The discriminative ability of the network at the five-second segment level is above chance level. Surprisingly, we noticed an outlier rat (rat #4) whose ROC curve is smaller than 0.5 which means that the network completely confused the segments from one phase

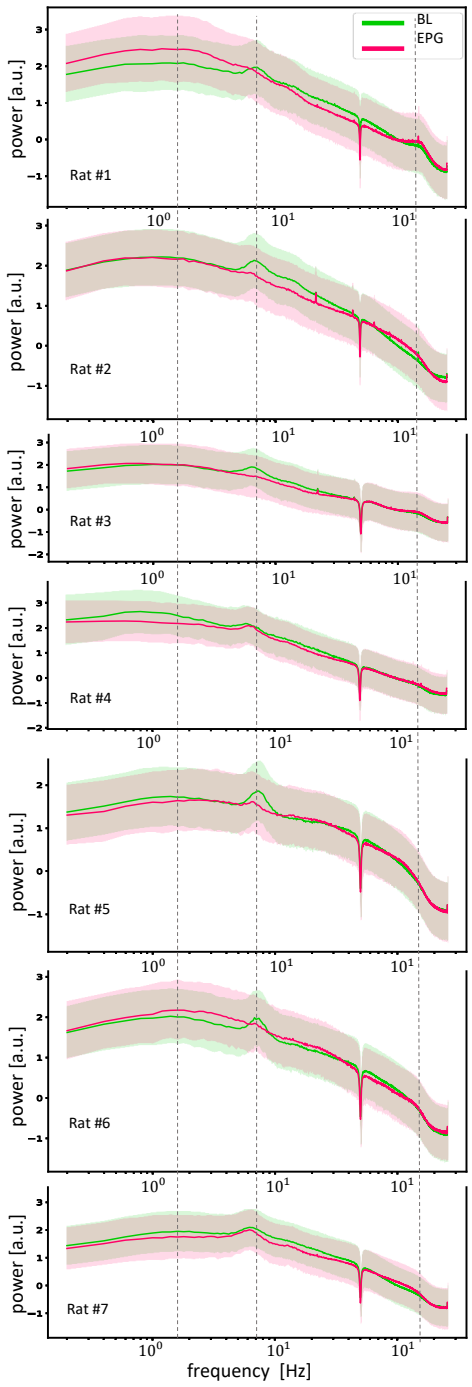


Fig. 7. Mean signal of the very certain samples of the two classes during testing from all LOO trials. Dashed lines are at 3 Hz, 7 Hz and 150 Hz. The shaded areas represent the standard deviation.

with the opposite phase. We refereed back to our recordings and so far we could not find any explanation why this rat is an outlier. Figure 5 shows the ROC curves without the outlier rat. We can see that in most of the cases our proposed method achieves significantly above-chance-level performance and the aggregation boosts the discriminative ability showing

TABLE II
PERFORMANCE **WITHOUT PREDICTION AGGREGATION**. VALUES ARE GIVEN WITH AND WITHOUT THE OUTLIER RAT #4

Tasks	SEN	SPE	AUC
w/ outlier rat	0.56 ± 0.2	0.58 ± 0.2	0.58 ± 0.2
w/o outlier rat	$\mathbf{0.62} \pm 0.2$	$\mathbf{0.62} \pm 0.2$	$\mathbf{0.64} \pm 0.1$

an increase of 0.26 in the averaged AUC value. The detailed specificity, sensitivity and AUC values with and without rat #4 in the two scenarios are given in Tab. III-D and Tab. III-D, respectively. The aggregation over one hour boosts performance.

Figure 6 shows the distribution of the scores collected from all the test trials with and without prediction aggregation. The score is defined as the probability of being an EPG segment. Ideally, true EPG segments would be scored close to one, and BL segments will be scored close to zero. The figure shows that in both scenarios, the distributions are significantly different from each other with an average p-value of $\ll 10^{-50}$ in a one-way ANOVA test for same population mean and average p-value of $\ll 10^{-20}$ in the Wilcoxon Rank Sum test to compare the two continuous distributions. Since the number of samples is sufficiently large, the p-values are driven to a very small value. To measure the sizes of differences between two groups, we also computed averaged Cohen's *d* effect size [19], which represents how far away the two population means are in the unit of standard deviation. An effect size that is smaller than 0.2 is considered a "small" difference, and a value that is bigger than 0.8 is considered a "big" difference [20]. In our experiment, the effect sizes with and without pooling are 0.31 ± 0.51 and 1.12 ± 1.81, respectively. Notably, from the individual violin plots we can see that there is still a certain overlap between BL and EPG segments, i.e., in BL period there are a certain amount of segments classified as EPG signals and vice versa.

We also plot the average of all the samples that the network is very sure about (the certainty is over 99.9%) during each LOO test trial. In Fig. 7 we show the averaged BL samples as well as the averaged EPG samples in each trial. These plots suggest that the network has a high confidence of EPG samples that have a high power around 3 Hz, a low power around 7 Hz, and slightly increased power around 150 Hz compare to BL samples. These findings are consistent with other works, e.g., Milikovsky et al. observed a decreased theta power in the epileptogeneic zone [8]. Jalilifar et al. [10] showed an increase of delta and a decrease of theta power in an epilepsy model based on kindling. Li et al. showed that the rate of hippocampal high frequency oscillations is increased in epileptogenesis [11].

IV. DISCUSSION

We explored the possibility to distinguish baseline (BL) and epileptogenesis (EPG) phases in a rodent epilepsy model with FFT transformed EEG data. We proposed a deep neural network framework for classification and a prediction aggregation process. We collected our data from a well established

TABLE III
PERFORMANCE **WITH ONE HOUR OF PREDICTION AGGREGATION**. VALUES ARE GIVEN WITH AND WITHOUT THE OUTLIER RAT #4

	SEN	SPE	AUC
w/ outlier rat	0.72 ± 0.28	0.72 ± 0.27	0.77 ± 0.31
w/o outlier rat	$\mathbf{0.83} \pm 0.1$	$\mathbf{0.83} \pm 0.08$	$\mathbf{0.90} \pm 0.08$

rodent model where epilepsy is introduced through perforant path stimulation (PPS). To test the generalization ability of our approach to unseen data collected from an unseen rat, we adopted a leave-one-out (LOO) cross validation scheme. The LOO test trials showed that our model generalized well to data from unseen rats with one exception. The reason for this "outlier rat" are presently unclear. The prediction aggregation over a longer period of time yielded better results in sensitivity, specificity and AUC. The inspection of the distribution of scores assigned to BL and EPG samples showed that there is a distribution shift in BL and EPG phases, i.e., there were a large number of BL samples with close-to-zero scores and a small number with high scores and vice versa. Further analysis of the samples that the network was very certain about (the certainty ≥ 0.999), we found that compared to the BL phase, the EPG phase recordings show a high power around 3 Hz, a substantial decrease of power around 7 Hz, and a slight increase of power around 150 Hz. These findings are consistent with the conclusions from other studies [8]–[11]. Overall, the predictions from our proposed network using the prediction aggregation method for one hour recordings are quite promising. Future work should address if similar results can also be obtained in human subjects with non-invasive recordings.

ACKNOWLEDGMENT

This work is supported by the China Scholarship Council (No. [2016]3100), the LOEWE Center for Personalized Translational Epilepsy Research (CePTER), and the Johanna Quandt Foundation.

REFERENCES

[1] S. L. Moshé, E. Perucca, P. Ryvlin, and T. Tomson, "Epilepsy: new advances," *The Lancet*, vol. 385, no. 9971, pp. 884–898, 2015.

[2] W. Löscher, "The holy grail of epilepsy prevention: preclinical approaches to antiepileptogenic treatments," *Neuropharmacology*, 2019.

[3] P. Kwan and M. J. Brodie, "Early identification of refractory epilepsy," *New England Journal of Medicine*, vol. 342, no. 5, pp. 314–319, 2000.

[4] K. Tzimourta, A. Tzallas, N. Giannakeas, L. Astrakas, D. Tsalikakis, and M. Tsipouras, "Epileptic seizures classification based on long-term eeg signal wavelet analysis," in *Precision medicine powered by pHealth and connected health*, pp. 165–169, Springer, 2018.

[5] S. M. Usman and A. Hassan, "Efficient prediction and classification of epileptic seizures using eeg data based on univariate linear features.," *JCP*, vol. 13, no. 6, pp. 616–621, 2018.

[6] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, p. 65, 2019.

[7] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, *et al.*, "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[8] D. Milikovsky, I. Weissberg, L. Kamintsky, K. Lippman, O. Schefenbauer, F. Frigerio, M. Rizzi, L. Sheintuch, D. Zelig, J. Ofer, *et al.*, "Theta rhythm alterations–a novel predictive biomarker of epilepsy," *Journal of the Neurological Sciences*, vol. 381, p. 86, 2017.

[9] L. Chauviere, N. Rafrafi, C. Thinus-Blanc, F. Bartolomei, M. Esclapez, and C. Bernard, "Early deficits in spatial memory and theta rhythm in experimental temporal lobe epilepsy," *Journal of Neuroscience*, vol. 29, no. 17, pp. 5402–5410, 2009.

[10] M. Jalilifar, A. Yadollahpour, A. A. Moazedi, and Z. Ghotbeddin, "Quantitative analysis of the antiepileptogenic effects of low frequency stimulation applied prior or after kindling stimulation in rats," *Frontiers in physiology*, vol. 9, 2018.

[11] L. Li, M. Patel, J. Almajano, J. Engel Jr, and A. Bragin, "Extrahippocampal high-frequency oscillations during epileptogenesis," *Epilepsia*, vol. 59, no. 4, pp. e51–e55, 2018.

[12] B. A. Norwood, S. Bauer, S. Wegner, H. M. Hamer, W. H. Oertel, R. S. Sloviter, and F. Rosenow, "Electrical stimulation-induced seizures in rats: a "dose-response" study on resultant neurodegeneration," *Epilepsia*, vol. 52, no. 9, pp. e109–e112, 2011.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[14] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, p. 3, 2013.

[17] P. Smyth and D. Wolpert, "Linearly combining density estimators via stacking," *Machine Learning*, vol. 36, no. 1-2, pp. 59–83, 1999.

[18] Y. Yao, A. Vehtari, D. Simpson, A. Gelman, *et al.*, "Using stacking to average bayesian predictive distributions (with discussion)," *Bayesian Analysis*, vol. 13, no. 3, pp. 917–1003, 2018.

[19] M. E. Rice and G. T. Harris, "Comparing effect sizes in follow-up studies: Roc area, cohen's d, and r," *Law and human behavior*, vol. 29, no. 5, pp. 615–620, 2005.

[20] S. S. Sawilowsky, "New effect size rules of thumb," *Journal of Modern Applied Statistical Methods*, vol. 8, no. 2, p. 26, 2009.

Towards Early Diagnosis of Epilepsy from EEG Data

Diyuan Lu[1,2,3] ELU@FIAS.UNI-FRANKFURT.DE

Sebastian Bauer[3,4] SEBASTIAN.BAUER@KGU.DE

Valentin Neubert[5] VALENTIN.NEUBERT@UNI-ROSTOCK.DE

Lara Sophie Costard[6] LARACOSTARD@RCSI.COM

Felix Rosenow[3,4] ROSENOW@MED.UNI-FRANKFURT.DE

Jochen Triesch[1,2,3] TRIESCH@FIAS.UNI-FRANKFURT.DE

[1]**Frankfurt Institute for Advanced Studies (FIAS), Frankfurt am Main, Germany**
[2]**Goethe University Frankfurt, Frankfurt am Main, Germany**
[3]**Center for Personalized Translational Epilepsy Research (CePTER), Frankfurt am Main, Germany**
[4]**Epilepsy Center Frankfurt Rhein-Main, University Hospital Goethe-University, Frankfurt am Main, Germany**
[5]**Oscar Langendorff Institute of Physiology, Rostock University Medical Center, Rostock, Germany**
[6]**Tissue Engineering Research Group, Royal College of Surgeons Ireland, Dublin, Ireland**

Abstract

Epilepsy is one of the most common neurological disorders, affecting about 1% of the population at all ages. Detecting the development of epilepsy, i.e., epileptogenesis (EPG), before any seizures occur could allow for early interventions and potentially more effective treatments. Here, we investigate if modern machine learning (ML) techniques can detect EPG from intra-cranial electroencephalography (EEG) recordings prior to the occurrence of any seizures by a time frame of days or even weeks. We study a common form of epilepsy called mesial temporal lobe epilepsy (mTLE). Specifically, we use a rodent mTLE model where EPG is triggered by electrical stimulation of the brain, which induces hippocampal damages that resemble those in human patients. We propose a ML framework for EPG identification, which combines a deep convolutional neural network (CNN) with a prediction aggregation method to obtain the final classification decision. Specifically, the neural network is trained to distinguish five second segments of EEG recordings taken from either the pre-stimulation period or the post-stimulation period. Due to the gradual development of epilepsy, there is enormous overlap of the EEG patterns before and after the stimulation.

Hence, a prediction aggregation process is introduced, which pools predictions over a longer period. By aggregating predictions over one hour, our approach achieves an area under the curve (AUC) of 0.99 on the EPG detection task. This demonstrates the potential of ML for EPG prediction from EEG recordings.

1. Introduction

Identifying patients at high risk of developing epilepsy (epileptogenesis) is of great importance to allow early medical intervention and improve the effectiveness of anti-epileptogenic treatments. In many acquired epilepsy cases, there is a latent period between the brain injury and the onset of spontaneous recurring seizures. During this latent period, affected brain tissue is thought to transform such that it eventually can generate spontaneous seizures (Pitkänen and Engel, 2014). Over 30% of the patients will be pharmaco-resistant and continue to suffer from recurring seizures despite intake of medications (Kwan and Brodie, 2000). The more seizure episodes have occurred before the first clinical visit, the less effective of the treatment will be (Kwan and Brodie, 2000). Hence, identifying the presence of EPG before the epilepsy is fully established would be of great importance. However, the process of EPG is still not fully understood (Pitkänen et al., 2016). The precise time of onset of the brain being epileptogenic is untraceable (Pitkänen and Engel, 2014). However, it is safe to say that any anti-epileptogenic or disease-modifying therapies should be administered as early as possible (Löscher, 2019). Thus, discovering prominent features of EPG could facilitate early diagnosis and open the door for early interventions (Moshé et al., 2015).

Electroencephalography (EEG) is a common tool in the clinic due to its non-invasive and easy-to-deploy properties. However, detecting EPG from EEG data is challenging. Two reasons are the complexity of the mechanisms of EPG and the immense cross-subject variability, which result in different phenotypes of EEG signals. This makes reliable interpretation of EEG signals from previously unseen individuals difficult.

Some works have attempted to identify electrophysiological biomarkers of EPG based on various hand-selected features (Bentes et al., 2018; Rizzi et al., 2019; Milikovsky et al., 2017; Bragin et al., 2004, 2016). However, a manual selection of features may be biased and overlook useful information. Recently, fueled by advances in ML, impressive results have been achieved in a variety of domains by training on raw data and letting the learning algorithm identify useful features automatically. Such approaches can even outperform human experts (Hannun et al., 2019; Haenssle et al., 2018; Sarker et al., 2018).

Here, we recorded intracranial EEG signals from a rodent model of mesial temporal lobe epilepsy with hippocampal sclerosis (mTLE-HS) (Costard et al., 2019). In this model, epilepsy was induced by electrical perforant pathway stimulation (PPS) through depth electrodes. Continuous EEG recordings were obtained from the hilus of the dentate gyrus after the implantation of the electrode until the occurrence of the first spontaneous seizure (FSS). The EEG recordings were divided into two classes depending on the time of recording relative to the PPS stimulation. The samples recorded before the epilepsy-triggering PPS define the *baseline* (BL) class. The samples recorded after the PPS, but before occurrence of the FSS form the *epileptogenesis* (EPG) class. In the following, we propose a deep learning

(DL) framework to classify EPG vs. BL by training on raw EEG data in an end-to-end fashion.

To tackle the problem that a large portion of normal brain EEG patterns are also present in the EPG phase, we propose a prediction aggregation method where predictions from a longer time interval (e.g. one hour) are pooled together through a linear aggregation. We assume that the "EPG-typical" signals should be more frequent during the EPG phase compared to the baseline phase. This difference becomes apparent through the aggregation method. Specifically, we make the following contributions:

- We present the first attempt to identify the process of EPG with a deep neural network (DNN) trained on EEG time series data. This is a radical departure from the conventional (and hitherto not very successful) approach of attempting to predict individual seizures when the disease has already established itself.
- We propose a framework for EPG identification using massive amounts of EEG data from chronic recordings to maximally exploit the DNN's learning ability and minimize human effort in data labeling and feature engineering.
- We use a prediction aggregation method and demonstrate that it achieves high fidelity EPG detection in a rodent model.

Generalizable Insights about ML in the Context of Healthcare

Massive expert annotations are expensive and therefore often scarce in medical contexts. This poses tremendous difficulties for the application of ML. When large amounts of data can be collected but labelling by experts is infeasible, turning to a form of "cheap" labelling can be a way-out. In our case, detailed expert annotations are absent but the EEG signals are recorded continuously (24/7), which yields a large quantity of training data. We define the labels exclusively according to the relative time of the recording with respect to the PPS. This kind of label is cheap and easy to obtain but less informative, since in the EPG period large amounts of normal brain activity are still present, i.e., the data from the two classes are largely overlapping. To deal with this large overlap, we propose a prediction aggregation process to pool decisions over a long time window. We show here that even in the complete absence of expert annotations of specific events showing "EPG-typical" brain activity, the large data set in combination with the "cheap" labels allow us to build a powerful classification system. We suggest that many other medical problems where the application of ML is currently infeasible due to lack of detailed expert annotations could be tackled using similar methods. More generally, our approach of massive data collection to identify the earliest signatures of a developing disease may enable early diagnosis and intervention across a wide range of medical contexts.

2. Related Work

EEG Analysis with Deep Learning Modern ML techniques allow an end-to-end learning approach to the analysis of EEG data rather than relying on specific handcrafted features. In particular, DNNs have been applied to either frequency representations (Lu et al.,

2019; Thodoroff et al., 2016) or directly to raw EEG data in the time domain (Kiral-Kornek et al., 2018; Biswal et al., 2019; Avcu et al., 2019; Farahat et al., 2019; Bi and Wang, 2019). They have achieved promising results in seizure detection, seizure prediction, or even other neurological disorders such as Alzheimer's disease and Autism classification. For example, Zhou et al. (2018) compared the performance of a CNN on the EEG signal classification problem with time-domain and frequency-domain input and concluded that frequency-domain signals have greater potential for the task. Kiral-Kornek et al. (2018) demonstrated an accurate, automated patient-specific seizure prediction approach with a DNN trained on intracranial EEG data. Biswal et al. (2019) applied stacked CNNs and recurrent neural networks (RNNs) to extract temporal shift invariant features from EEG data. These features are used to classify multiple key EEG phenotypes. Avcu et al. (2019) developed an end-to-end solution for seizure onset detection. Bi and Wang (2019) applied a convolutional deep Boltzmann machine with EEG data in early diagnosis of Alzheimer's disease. Thodoroff et al. (2016) applied a deep RNN with a CNN to perform automated patient specific seizure detection with scalp EEG. A deep CNN was applied for EEG signal decoding during human decision making and demonstrates promising results (Farahat et al., 2019). These studies demonstrated the application of DL for EEG analysis.

Here, we want to emphasize the fundamental difference between seizure prediction and our task. The goal of epileptic seizure prediction is to predict the onset of individual seizures in an epileptic brain that already generates spontaneous seizures. The goal is typically to predict individual seizures several minutes before their occurrence, so the patient can be warned about the imminent seizure and take precautions. In contrast, we aim to detect if a brain is on its way to develop an epilepsy *before* the FSS has occurred, i.e. before an epilepsy is manifest. If this could be done several days or weeks before the FSS, this would allow for interventions that could slow down or even prevent the development of the disease, before spontaneous seizures occur.

EPG Biomarkers in EEG There have been several previous studies on biomarker discovery for identifying EPG. Bragin et al. (2004) found that the occurrence of high-frequency-oscillations (HFOs) is a strong indicator of future recurrent spontaneous seizures and the sooner HFOs occur, the shorter the EPG period will be. Andrade et al. (2017) found that a duration reduction of sleep spindles at the transition from stage III to rapid-eye-movement sleep indicates potential post-traumatic epilepsy in a lateral fluid-percussion rat model. In humans, it was shown that over 90% of the HFO area overlapped with the seizure onset zone for six patients (Burnos et al., 2014). Milikovsky et al. (2017) revealed that the dynamics of the theta band could predict future post-injury epilepsy and the seizure onset and thus could serve as a diagnostic biomarker for EPG. Lu et al. (2019) demonstrated that an increased delta band power, a decrease of theta band power as well as an increase of high gamma band power were correlated with the presence of EPG in a rat mesial temporal lobe epilepsy model. Rizzi et al. (2019) recently showed using concepts from nonlinear dynamics, that a reduction of the dimensionality of EEG/ECoG signals indicates the presence and the severity of EPG in three different rodent epilepsy models. Finally, Bentes et al. (2018) found that an asymmetry in background EEG signals and interictal epileptiform discharges can independently predict post-stroke epilepsy in a clinical study. However, so far a DL-based approach to EPG biomarker discovery in an end-to-end fashion has not yet been attempted.

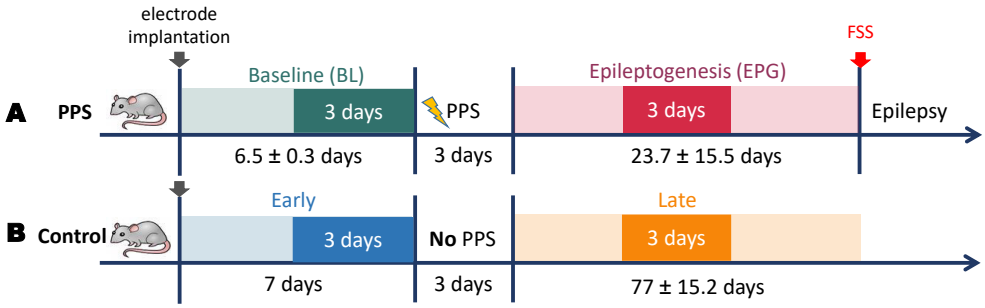


Figure 1: Schematic of the timeline of the experiment. A. time line for PPS-stimulated rats. B: time line for control rats. PPS: perforant pathway stimulation. FSS: first spontaneous seizure. The mean and standard deviation of the duration of EPG in the PPS group is 23.7 ± 15.5 days (min. 10 days, max. 56 days).

3. Methods

3.1. Dataset

We used intracranial EEG data recorded continuously (24/7) by a depth-electrode from a rodent mTLE-HS model, where epilepsy is induced by electrical PPS, as described in detail in Costard et al. (2019). The stimulated rats developed epilepsy after an average EPG phase of four weeks (range one to eight weeks). The EPG phase ended with the FSS.

The rat model provides an opportunity to study the progression of epilepsy and to discover potential biomarkers of EPG in the EEG. In this study, we included seven PPS-treated rats with continuous wireless EEG recordings. We also included three control rats which had electrodes implanted but did not undergo PPS and did not develop epilepsy by the end of the recording (limited by the lifetime of battery of the wireless transmitter). The time-lines for the PPS group and the control group are shown in Fig. 1. We denoted two phases of interest from the continuous recording, i.e., baseline (BL) and epileptogenesis (EPG) in the PPS group. In our study, we selected the last three days from the BL phase and three days from the EPG phase, highlighted in the colored boxes, and assigned them the labels "0" and "1", respectively. We selected the 7th, 8th and 9th day of EPG for training for all rats. Reasons for this choice are 1) to maintain the maximum time distance to acute symptomatic seizures which can occur within the first 1-3 days after the PPS, and 2) the rat with the shortest EPG duration developed its FSS on the 10th day after PPS and we wanted to keep the time window from which we get the class "1" signals the same across all rats.

Preprocessing The sampling rate of the EEG recordings was 512 Hz. A band-pass filter between 0.5 - 160 Hz and a notch filter at 50 Hz were applied to the raw data. In our experimental setting, the recorded EEG signals were susceptible to electric interference, which resulted in extremely high amplitude outliers. To fix this problem, we applied a MATLAB function, i.e., `filloutliers` [1] with the configuration `method = 'pchip';`

1. https://www.mathworks.com/help/matlab/ref/filloutliers.html

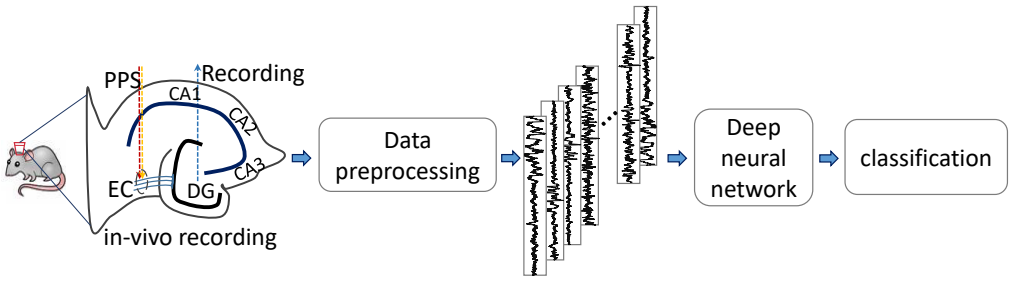


Figure 2: Workflow of our proposed framework. EC: entorhinal cortex, DG: dentate gyrus, CA: cornu ammonis, PPS: perforant pathway stimulation.

`movmethod = 'movmedian'; window = 50` to filter out these outliers. We obtained non-overlapping five-second segments from the continuous recordings. To clean up the data for training, those segments with more than 20% data loss due to weak wireless transmission were discarded. Then, those five-second segments were normalized via the z-score method from `scipy.stats.zscore` before being fed into the neural network. The workflow is shown in Fig 2.

Our proposed method consists of two parts: (a) a deep residual neural network and (b) a prediction aggregation process during the testing.

Residual convolutional neural network Our model is a DNN with 33 convolutional layers with residual connections and it is inspired by the work of Hannun et al. (2019). The network's structure is shown in Table 1. The concept of residual connections was first proposed by He et al. (2016a) for an image recognition task and has been widely used in a variety of tasks such as image segmentation (Huang et al., 2017; Lei and Todorovic, 2018; Liu et al., 2019), visual object detection (Mordan et al., 2018; Wang et al., 2019), and healthcare-related applications (Hannun et al., 2019; Sarker et al., 2018). The residual connection connects the pre-activation from one layer with the input of another previous layer in an additive fashion skipping several layers in between. Then, the non-linear activation is applied to the sum to compute the input for the next layer. The collection of the computations between one residual connection is termed a block (ResBlock). The output of the network is a softmax layer taking the flattened feature maps as input and outputting a probability distribution over the two possible classes.

Before we started our official classifier training, we performed a hyper-parameter exploration for our specific task with a small randomly selected data set. A drop-out rate of 0.25 yielded the best performance among the values 0.2, 0.25, 0.3, 0.5, and 0.65. The number of blocks that performed best was 15 among 5, 7, 11, and 15. A filter size of 32 worked the best among values of 3, 9, 11, 16, 32, and 64. We tried ReLU and leaky ReLU as the nonlinear activation function and no significant difference was observed, so we chose the ReLU activation for this work. A starting number of 16 filters yielded better results than 8 and 32. After the network hyper-parameter exploration, we fixed the choices for further experiments.

Table 1: The network structure used in our work. The **Config** column show the filter size (always 32) and the number of filters we use in each convolutional layer. The number of filters is increased every four blocks by a factor of 2. Every other block sub-samples its input by a factor of 2, indicated by the value of **stride**. Here, the batch size at the first dimension is omitted in the output shape column

Name	**Config**	**Stride**	**Factor i**	**Output shape**
Conv layer 0	$\begin{bmatrix} 32\times 1, & 16\times 2^i \end{bmatrix}$	1	0	[2560, 1, 16]
ResBlock 0	$\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$	1	0	[2560, 1, 16]
ResBlock 1	$\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$	2	0	[1280, 1, 16×2^i]
ResBlock 2	$\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$	1	0	[1280, 1, 16×2^i]
ResBlock 3	$\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$	2	0	[640, 1, 16×2^i]
ResBlock 4	$\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$	1	1	[640, 1, 16×2^i]
ResBlock (5,..., 8)	$\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$	(2, 1, 2, 1)	(1, 1, 1, 2)	[320, 1, 16×2^i]
ResBlock (9,..., 12)	$\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$	(2, 1, 2, 1)	(2, 2, 2, 3)	[80, 1, 16×2^i]
ResBlock (13, 14)	$\begin{bmatrix} 32\times 1, & 16\times 2^i \\ 32\times 1, & 16\times 2^i \end{bmatrix}$	(2, 1)	(3, 3)	[20, 1, 16×2^i]
Dense	2			[2]

We adopted the pre-activation design from He et al. (2016b). The convolutional layer had a filter width of 32. The number of filters increased by a factor of 2 in every four blocks starting from 16. The feature maps were down-sampled in every other block with a stride of 2. To keep the dimensionality compatible, the max-pooling branch shared the same stride value as in the second convolutional layer in each block.

Prediction aggregation We hypothesize that the EPG phase may be better characterized by a change of distribution of different waveforms rather than a specific waveform that can be identified in every individual segment. Therefore a reliable classification can only be achieved by pooling information from many data segments. Our method is inspired by Smyth and Wolpert (1999). For each segment, the network outputs how likely this segment is taken from each class. We linearly aggregate the predictions for multiple consecutive segments to obtain the final classification result.

Considering the data pairs, the EEG segments are $x_{(h,i)}$ and the associated labels are $y_{(h,i)}$ in one continuous hour h, where $i = 1, \ldots, N$ and N is the total number of the samples in this hour. The softmax output of these samples is given by $\hat{y}_{(h,i)} = f(x_{(h,i)}, \text{model})$ and it is in shape $[N, 2]$ where 2 is the number of classes in our supervised scheme. The aggregated

Table 2: Performance **without** (5 second) and **with** one hour of aggregation. Data are presented as mean ± standard deviation. SEN: sensitivity, SPE: specificity, AUC: area under the curve

Aggregation length	Task	SEN	SPE	AUC
5 second	Task A	0.73 ± 0.25	0.77 ± 0.17	0.86 ± 0.07
	Task B	0.57 ± 0.42	0.43 ± 0.42	0.50 ± 0.08
1 hour	Task A	$\mathbf{0.94} \pm 0.05$	$\mathbf{0.96} \pm 0.04$	$\mathbf{0.99} \pm 0.01$
	Task B	0.63 ± 0.45	0.37 ± 0.45	0.45 ± 0.06

prediction for hour h is given by $\hat{y}_h = \sum_{i=1}^{N} \hat{y}_{(h,i)} = \sum_{i=1}^{N} f(x_{(h,i)}, \text{model})$, and in shape of $[1, 2]$. In a final step, we normalize $\hat{y}_h$ along the column axis. The resulting number is interpreted as a class probability and used to compute corresponding performance metrics.

Training procedure We applied leave-one-out (LOO) cross validation to test the generalization ability of our approach for both the PPS group and the control group. Specifically, in each fold the data from one rat were completely withheld as the test set, and the data from the other six rats form the training and the validation sets. For training and validation, we randomly selected 25 hours from each phase and from each rat and applied a train-validation-split of 9:1. The choice of 25 hours represents a trade-off between computation cost and performance, chosen empirically. We tried training with the whole three-day recordings, and the computation time was increased by a factor of three without obvious performance improvement. After the network was trained, we tested it with the data from the previously held-out rat. The procedure was repeated seven times in the PPS group and three times in the control group and results were averaged for each group.

4. Experiments and Results

4.1. Experiment Design

To evaluate our methods ability to identify EPG, we designed two tasks: Task A is designed to classify BL vs. EPG signals in PPS rats as shown in Fig 1A. Task B is a control designed to classify signals recorded in the early and late implantation phases in the set of control rats as shown in Fig 1B.

Task A: BL vs. EPG classification in PPS rats This is our main task in which we want to distinguish EEG signals from BL and EPG phases. In this task, we applied seven-fold LOO cross validation with the data from the seven PPS-stimulated rats.

Task B: *early* vs. *late* classification in control rats In this control task we want to rule out the possibility that differences between BL and EPG in Task A could be simply due to systematic changes in the tissue after electrode implantation that have nothing to do with the EPG triggered by PPS. Therefore, we study control rats that do not undergo PPS (see Fig 1B) and analyze if there are systematic differences between the EEG signals

recorded from the early and late implantation phases. We applied a three-fold LOO cross validation scheme with the same network configuration as in Task A.

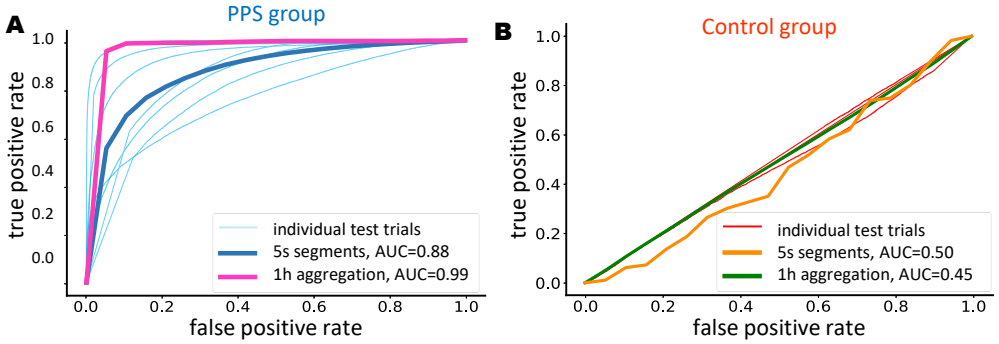


Figure 3: Receiver operating characteristic (ROC) curves. **A**: **The PPS group** (seven rats). Individual ROC curves from all LOO test trials (thin light blue), the average ROC curve without prediction aggregation (thick blue) and the average ROC curve with aggregation in a continuous stretch of one hour (thick pink). **B**: **The control group**. Individual ROC curves (thin light orange), the average ROC curve without aggregation (thick orange) and the average ROC curve with aggregation over one hour (thick green) from all LOO test trials in the control group. AUC: area under the curve. PPS: perforant pathway stimulation. LOO: leave-one-out

4.2. Results

4.2.1. ROC Analysis

The average ROC curves of all the leave-one-out test trials in each task are shown in Fig 3. The AUC values are computed in two scenarios: a) each five second segment is viewed independently and the AUC is calculated based on the prediction of all the five second segments, b) the predictions of multiple consecutive five second segments are aggregated together through a linear stacking. In Fig 3A, we show the ROC curves in individual LOO test trials, and the averaged ROC curves with and without prediction aggregation. Our method could discern signals from both phases with an average AUC under the ROC curve of 0.88. It suggests that the neural network has learned features that are informative for the correct classification. With the proposed prediction aggregation over one hour, the average AUC achieves 0.99, which shows that the proposed approach can reliably discern EEG signals from the BL and the EPG phase. In contrast, for the control group, the *early* vs. *late* phase classification, the network does not show clear discriminative ability. The average AUCs from all the test trials with and without the prediction aggregation are 0.50 and 0.45, respectively. The detailed performance measurements such as sensitivity (SEN) $= \frac{TP}{TP+FN}$, specificity (SPE) $= \frac{TN}{TN+FP}$ and the AUC are shown in Table. 2, where *TP, TN, FP, FN* denote true positive, true negative, false positive and false negative, respectively.

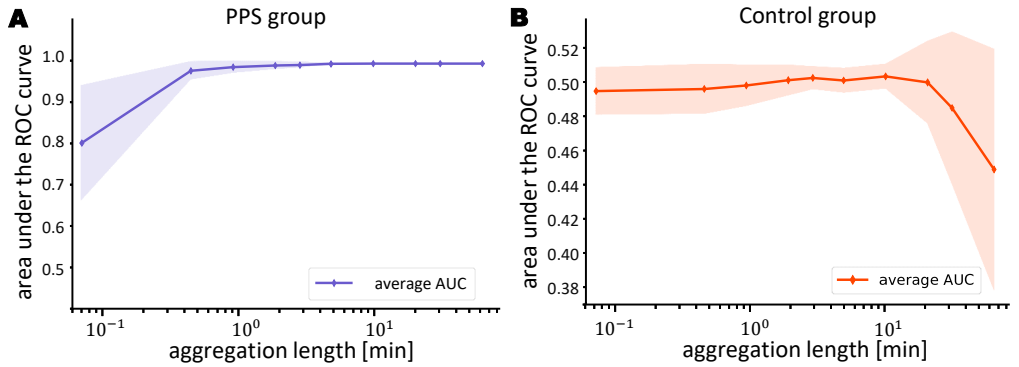


Figure 4: The average AUC over all leave-one-out test trials as a function of the aggregation length for the two groups. The shaded area represents one standard deviation.

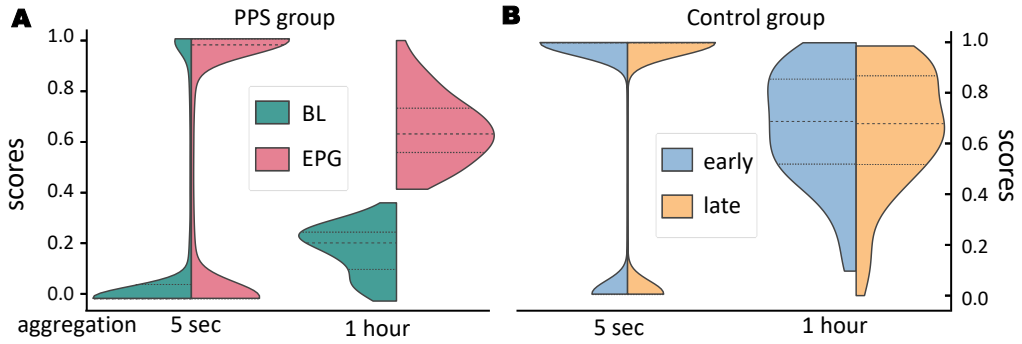


Figure 5: Example distributions of scores from both classes. **A**: **PPS group**. **(left)** without aggregation. The mean and variance of the two distributions, i.e., from all BL segments and all EPG segments, are different but overlapping. **(right)** with one hour of aggregation. **B**: **Control group**. **(left)** without aggregation. **(right)** with one hour of aggregation. BL: baseline. EPG: epileptogenesis

4.2.2. Aggregation Effect

To further investigate the effect of aggregation, we computed the AUC value in each test trial with various intervals, i.e., five seconds, 30 seconds, one, two, five, ten, 20, 30, 60 minutes. The average AUC across all the test trials in the PPS group as a function of the aggregation lengths is shown in Fig 4A. It shows a clear trend of an increasing AUC and a decrease of standard deviation with a longer aggregation length. Thus, the prediction aggregation from multiple consecutive segments is essential for a strong performance in the PPS group. In contrast, in the control group, the aggregation not only did not help increase but reduced the average AUC, as depicted in Fig 4B.

We also tested if the seven neural networks trained on the PPS group would discriminate the *early* and *late* phase EEG patterns from the control animals. If so, this would suggest

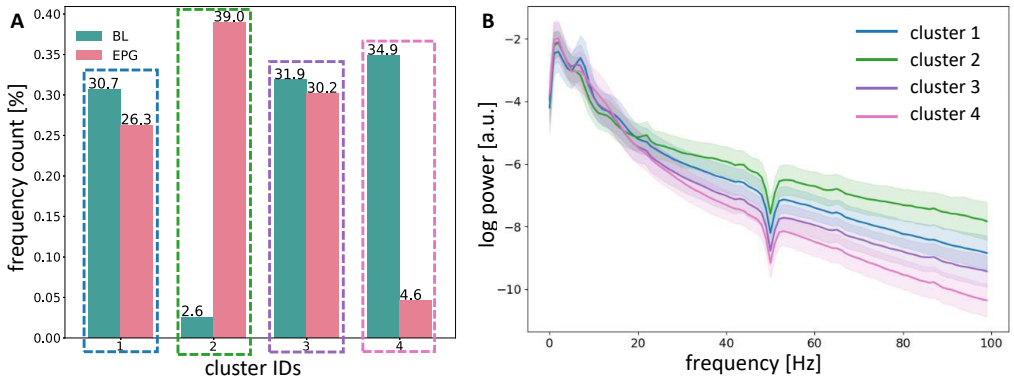


Figure 6: Clustering of high EPG score examples. **A**: percentage count of each class in each cluster. **B**: mean spectra of each cluster. The shaded area represents the standard deviation.

that these networks learn to discover changes in the EEG patterns across time that are triggered by the surgical procedure but are independent of the PPS and the ensuing EPG. However, we found that these networks could not discriminate *early* and *late* EEG patterns from the control group (mean AUC = 0.53, std. dev. = 0.12) and over 82% of all test samples from both *early* and *late* phases are classified as BL. This is additional evidence that the networks have learned to detect changes in EEG patterns that are induced by the PPS.

To visualize how exactly the prediction aggregation improves the discriminative ability of the model, we compute the distribution of scores assigned by the network to all test segments. Notably, the **score** is defined as the softmax output of the segment being EPG. Ideally, scores for BL segments should be close to zero, and EPG segments should have close-to-one scores. For simplicity, we only show the distributions of one representative LOO test trial from each group, as presented in Fig 5. The difference of the distributions within the same aggregation length is evaluated with the ANOVA test and the Wilcoxon rank sum test. In Fig. 5A, the distributions are significantly different in both cases for this rat (the ANOVA test, p-value $\leq 10^{-25}$, the Wilcoxon rank sum test, p-value $\leq 10^{-17}$). Results for other PPS rats are similar (not shown). To measure the sizes of differences between two distributions within the same aggregation length, we also computed Cohen's d effect size (Rice and Harris, 2005). In the two examples shown, $d = 0.94$ and 2.91, respectively. Average d values for the whole PPS-stimulated group with and without aggregation are 0.85 and 1.24, respectively. Cohen suggested that an effect size absolute value over 0.8 is considered large. Notably, there is still a considerable overlap between BL and EPG segments, i.e., in the BL period there are a certain number of segments classified as EPG and vice versa. When we aggregate over one hour, the effect of the distribution shift is magnified. In contrast, in the control group, the distributions of scores from one representative test trial with and without aggregation, are shown in Fig 5B, are not significantly different (the ANOVA test, p-values ≥ 0.5, the Wilcoxon rank sum test, p-values ≥ 0.4) with an effect size $d = 0.004$ and 0.012, respectively. The other two LOO test trials in the control group exhibit the same pattern.

4.2.3. K-Means Clustering Analysis

In order to obtain a better understanding of the characteristics of the learned features, we conducted k-means clustering analysis on very certain samples collected from LOO test-trials. Here, a certain sample is defined as one whose softmax probability is larger than a threshold (set to 0.999). The k-means clustering analysis is based on the Euclidean distance between two samples power spectra. Specifically, we cluster the log-power spectrum of examples into four clusters, where the number of four is determined by the elbow-theory (Kodinariya and Makwana, 2013). From the frequency count plot, see Fig. 6A, we can see that the majority of the cluster No. 2 stems from the EPG class and that of the cluster No. 4 is from the BL class. From the mean spectra of each cluster, we can see that the EPG-dominant cluster has higher power in the frequency range over 20 Hz to 100 Hz. Specially, in this cluster, there is strong power around 22 Hz and its harmonics. On the other hand, the mean power spectrum of the BL-dominant cluster, cluster No. 4, has a faster decay towards higher frequencies.

5. Discussion

In recent years, ML could capitalize on the availability of big medical data sets. However, acquiring expert annotations for such data is impractical in many applications, representing a challenge for ML approaches. Here, we have tried to answer the question if an emerging epilepsy might be detectable from EEG signals even before the first seizure occurs. For this, we have used a rodent model of epilepsy (Costard et al., 2019), where epileptogenesis (EPG) is triggered through PPS. While massive amounts of training data are available from the BL (pre-stimulation) and the EPG (post-stimulation) periods, these data are only labeled by their time of recording. On the one hand, there might be large amounts of EPG-like signals present in the BL phase because there is brain injury involved in implanting the electrode. On the other hand, normal brain activities are still present in the EPG phase. Thus, we can expect short segments of EEG recordings to be often indistinguishable. A reliable classification requires pooling data over longer time windows. To achieve this, we have proposed a DNN approach with a prediction aggregation method. Our method is trained in an end-to-end fashion on five second segments and we have observed massive performance gains when aggregating predictions over one hour (improvements of 21%, 19%, and 13% in SEN, SPE, and AUC, respectively). Therefore, we have demonstrated a viable method for automatically predicting epilepsy from EEG recordings prior to the first epileptic seizure. This opens the door for early interventions to modify or even arrest the progression of the disease (Löscher, 2019). Furthermore, EEG patterns that the network has identified as being predictive of EPG may point towards new biomarkers of the disease. As a plausible alternative approach to our network architecture, a recurrent neural network (RNN) could be considered. However, our preliminary investigations have shown that RNN training requires more structure exploration and hyper-parameter search and our results leave little room for improvement on the data set presented here.

Limitations From the perspective of practical utility, a good biomarker for identifying EPG in a clinical setting should be noninvasive. In contrast, the data in our study were recorded with a depth electrode, which has a much higher signal-to-noise-ratio compared

to surface EEG recordings. For training a similar model to predict EPG in humans, the collection of surface EEG data from human patients would be necessary. As an immediate next step, we plan to extend our results to a group of human patients, who will undergo EEG (surface or intracranial) recording in the hospital after suffering a brain injury but before epilepsy is manifest. With sufficient training data from these and non-epileptic patients, we could envision a machine-learning-assisted diagnostic tool for the early detection of a developing epilepsy in human patients.

References

Pedro Andrade, Jari Nissinen, and Asla Pitkänen. Generalized seizures after experimental traumatic brain injury occur at the transition from slow-wave to rapid eye movement sleep. *Journal of neurotrauma*, 34(7):1482–1487, 2017.

Mustafa Talha Avcu, Zhuo Zhang, and Derrick Wei Shih Chan. Seizure Detection Using Least Eeg Channels by Deep Convolutional Neural Network. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1120–1124. IEEE, 2019.

Carla Bentes, Hugo Martins, Ana Rita Peralta, Carlos Morgado, Carlos Casimiro, Ana Catarina Franco, Ana Catarina Fonseca, Ruth Geraldes, Patrícia Canhão, Teresa Pinho e Melo, et al. Early EEG predicts poststroke epilepsy. *Epilepsia open*, 3(2):203–212, 2018.

Xiaojun Bi and Haibo Wang. Early Alzheimers disease diagnosis based on EEG spectral images using deep learning. *Neural Networks*, 114:119–135, 2019.

Siddharth Biswal, Cao Xiao, M Brandon Westover, and Jimeng Sun. EEGtoText: Learning to Write Medical Reports from EEG Recordings. In *Machine Learning for Healthcare Conference*, pages 513–531, 2019.

Anatol Bragin, Charles L Wilson, Joyel Almajano, Istvan Mody, and Jerome Engel Jr. High-frequency oscillations after status epilepticus: epileptogenesis and seizure genesis. *Epilepsia*, 45(9):1017–1023, 2004.

Anatol Bragin, Lin Li, Joyel Almajano, Catalina Alvarado-Rojas, Aylin Y Reid, Richard J Staba, and Jerome Engel Jr. Pathologic electrographic changes after experimental traumatic brain injury. *Epilepsia*, 57(5):735–745, 2016.

Sergey Burnos, Peter Hilfiker, Oguzkan Sürücü, Felix Scholkmann, Niklaus Krayenbühl, Thomas Grunwald, and Johannes Sarnthein. Human intracranial high frequency oscillations (HFOs) detected by automatic time-frequency analysis. *PloS one*, 9(4), 2014.

Lara S Costard, Valentin Neubert, Morten T Venø, Junyi Su, Jørgen Kjems, Niamh MC Connolly, Jochen HM Prehn, Gerhard Schratt, David C Henshall, Felix Rosenow, et al. Electrical stimulation of the ventral hippocampal commissure delays experimental epilepsy and is associated with altered microrna expression. *Brain Stimulation*, 12(6): 1390–1401, 2019.

Amr Farahat, Christoph Reichert, Catherine M Sweeney-Reed, and Hermann Hinrichs. Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization. *Journal of neural engineering*, 16(6):066010, 2019.

Holger A Haenssle, Christine Fink, R Schneiderbauer, Ferdinand Toberer, Timo Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, L Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8): 1836–1842, 2018.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65, 2019.

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, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016b.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Isabell Kiral-Kornek, Subhrajit Roy, Ewan Nurse, Benjamin Mashford, Philippa Karoly, Thomas Carroll, Daniel Payne, Susmita Saha, Steven Baldassano, Terence O'Brien, et al. Epileptic seizure prediction using big data and deep learning: toward a mobile system. *EBioMedicine*, 27:103–111, 2018.

Trupti M Kodinariya and Prashant R Makwana. Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6):90–95, 2013.

Patrick Kwan and Martin J Brodie. Early identification of refractory epilepsy. *New England Journal of Medicine*, 342(5):314–319, 2000.

Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6742–6751, 2018.

Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 82–92, 2019.

Wolfgang Löscher. The holy grail of epilepsy prevention: Preclinical approaches to antiepileptogenic treatments. *Neuropharmacology*, 167:107605, 2019.

Diyuan Lu, Sebastian Bauer, Valentin Neubert, Lara Sophie Costard, Felix Rosenow, and Jochen Triesch. A Deep Residual Neural Network Based Framework for Epileptogenesis Detection in a Rodent Model with Single-Channel EEG Recordings. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6. IEEE, 2019.

Dan Z Milikovsky, Itai Weissberg, Lyn Kamintsky, Kristina Lippmann, Osnat Schefenbauer, Federica Frigerio, Massimo Rizzi, Liron Sheintuch, Daniel Zelig, Jonathan Ofer, et al. Electrocorticographic dynamics as a novel biomarker in five models of epileptogenesis. *Journal of Neuroscience*, 37(17):4450–4461, 2017.

Taylor Mordan, Nicolas Thome, Gilles Henaff, and Matthieu Cord. Revisiting multi-task learning with ROCK: a deep residual auxiliary block for visual detection. In *Advances in Neural Information Processing Systems*, pages 1310–1322, 2018.

Solomon L Moshé, Emilio Perucca, Philippe Ryvlin, and Torbjörn Tomson. Epilepsy: new advances. *The Lancet*, 385(9971):884–898, 2015.

Asla Pitkänen and Jerome Engel. Past and present definitions of epileptogenesis and its biomarkers. *Neurotherapeutics*, 11(2):231–241, 2014.

Asla Pitkänen, Wolfgang Löscher, Annamaria Vezzani, Albert J Becker, Michele Simonato, Katarzyna Lukasiuk, Olli Gröhn, Jens P Bankstahl, Alon Friedman, Eleonora Aronica, et al. Advances in the development of biomarkers for epilepsy. *The Lancet Neurology*, 15 (8):843–856, 2016.

Marnie E Rice and Grant T Harris. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and human behavior*, 29(5):615–620, 2005.

Massimo Rizzi, Claudia Brandt, Itai Weissberg, Dan Z Milikovsky, Alberto Pauletti, Gaetano Terrone, Alessia Salamone, Federica Frigerio, Wolfgang Löscher, Alon Friedman, et al. Changes of dimension of EEG/ECoG nonlinear dynamics predict epileptogenesis and therapy outcomes. *Neurobiology of disease*, 124:373–378, 2019.

Md Mostafa Kamal Sarker, Hatem A Rashwan, Farhan Akram, Syeda Furruka Banu, Adel Saleh, Vivek Kumar Singh, Forhad UH Chowdhury, Saddam Abdulwahab, Santiago Romani, Petia Radeva, et al. SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 21–29. Springer, 2018.

Padhraic Smyth and David Wolpert. Linearly combining density estimators via stacking. *Machine Learning*, 36(1-2):59–83, 1999.

Pierre Thodoroff, Joelle Pineau, and Andrew Lim. Learning robust features using deep learning for automatic seizure detection. In *Machine learning for healthcare conference*, pages 178–190, 2016.

Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7289–7298, 2019.

Mengni Zhou, Cheng Tian, Rui Cao, Bin Wang, Yan Niu, Ting Hu, Hao Guo, and Jie Xiang. Epileptic seizure detection based on EEG signals and CNN. *Frontiers in neuroinformatics*, 12(95), 2018.

Staging Epileptogenesis with Deep Neural Networks

Diyuan Lu
elu@fias.uni-frankfurt.de
Frankfurt Institute for Advanced Studies (FIAS)
Frankfurt am Main, Germany

Sebastian Bauer
Sebastian.Bauer@kgu.de
Epilepsy Center Frankfurt
Rhine-Main Neurocenter
Frankfurt am Main, Germany

Valentin Neubert
valentin.neubert@uni-rostock.de
Oscar-Langendorff-Institute for Physiology
Rostock, Germany

Lara Sophie Costard
laracostard@rcsi.com
Royal College of Surgeons Ireland
Dublin, Ireland

Felix Rosenow
rosenow@med.uni-frankfurt.de
Epilepsy Center Frankfurt
Rhine-Main Neurocenter
Frankfurt am Main, Germany

Jochen Triesch
triesch@fias.uni-frankfurt.de
Frankfurt Institute for Advanced Studies (FIAS)
Frankfurt am Main, Germany

ABSTRACT

Epilepsy is a common neurological disorder characterized by recurrent seizures accompanied by excessive synchronous brain activity. The process of structural and functional brain alterations leading to increased seizure susceptibility and eventually spontaneous seizures is called epileptogenesis (EPG) and can span months or even years. Detecting and monitoring the progression of EPG could allow for targeted early interventions that could slow down disease progression or even halt its development. Here, we propose an approach for staging EPG using deep neural networks and identify potential electroencephalography (EEG) biomarkers to distinguish different phases of EPG. Specifically, continuous intracranial EEG recordings were collected from a rodent model where epilepsy is induced by electrical perforant pathway stimulation (PPS). A deep neural network (DNN) is trained to distinguish EEG signals from before stimulation (baseline), shortly after the PPS and long after the PPS but before the first spontaneous seizure (FSS). Experimental results show that our proposed method can classify EEG signals from the three phases with an average area under the curve (AUC) of 0.93, 0.89, and 0.86. To the best of our knowledge, this represents the first successful attempt to stage EPG prior to the FSS using DNNs.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; Cross-validation; **Neural networks**; • **Applied computing** → **Bioinformatics**; • **Networks** → Network performance analysis.

KEYWORDS

epileptogenesis, deep neural network, machine learning, EEG, class activation map, feature visualization

BCB '20, September 21–24, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7964-9/20/09.
https://doi.org/10.1145/3388440.3412480

ACM Reference Format:
Diyuan Lu, Sebastian Bauer, Valentin Neubert, Lara Sophie Costard, Felix Rosenow, and Jochen Triesch. 2020. Staging Epileptogenesis with Deep Neural Networks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '20), September 21–24, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3388440.3412480

1 INTRODUCTION

Epilepsy is one of the most common and disruptive neurological disorders affecting about 1% of the world's population. It is characterized by recurrent unprovoked seizures and is accompanied by various co-morbidities such as migraine, depression, dementia, etc. [12]. Over 30% of the patients will eventually develop refractory epilepsy, defined as inadequate control of seizures by any medication [15]. In acquired epilepsy, an initial precipitating injury (IPI) such as stroke, traumatic brain injury or encephalitis leads to structural and functional remodelling of neuronal networks resulting in the occurrence of spontaneous seizures after a clinically silent latent period [20]. This remodelling process is termed epileptogenesis (EPG). Traditionally, epilepsy is diagnosed and treated after at least one unprovoked seizure, which indicates that the EPG has already progressed to a relatively advanced stage. This latent period can last months or even years. Treating high-risk patients at the early stage of EPG, or even customizing the treatment based on the severity of EPG could result in more effective disease-altering or even disease-arresting outcomes.

Pathomechnisms of EPG are not fully understood and its detection remains a major challenge. Studying early EPG in human patients is extremely difficult, simply because the epilepsy is typically only detected after the FSS. Therefore, work on early EPG is typically restricted to animal models [2]. Furthermore, early EPG can comprise a complex cascade of changes to the brain after the initial brain insult and this cascade may strongly depend on the type of brain insult. Changes can include, e.g., inflammatory reactions or blood-brain-barrier damage [9]. Some of these brain changes may be reflected in the EEG in the form of interictal epileptiform discharges (IEDs, including sharp-waves, spikes, spike-and-waves complex.), high-frequency oscillations, slowing or alteration of sleep spindles. Correspondingly, there have been attempts to identify suitable EEG biomarkers for EPG using a wide range of approaches [1, 3, 6, 17–19, 21]. However, a reliable staging

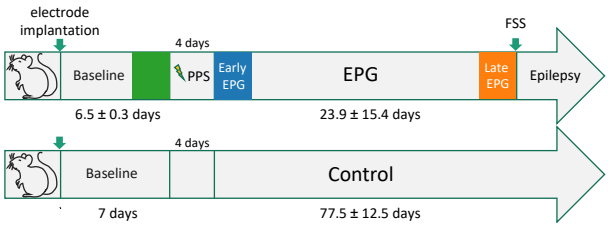


Figure 1: Timeline of the experiment. Shaded boxes indicate the different time periods where training and testing data are extracted. Upper row: PPS group. Lower row: control group (identical but without PPS). FSS: first spontaneous seizure. PPS: perforant pathway stimulation.

of EPG based on EEG measurements has not been demonstrated yet to the best of our knowledge.

Here, we use a rat epilepsy model, where EPG is induced by electrical perforant pathway stimulation (PPS) [8]. In previous work, we have shown that a DNN can be trained to distinguish EEG signals from baseline and EPG, i.e., before and after the PPS, with high specificity and sensitivity. Furthermore, we have demonstrated generalization to unseen rats [18]. Here we extend these results and present the first attempt to stage EPG using DNNs. In particular, we ask whether a DNN can also learn to distinguish early and late phases of EPG after the PPS but prior to the FSS, thereby allowing to estimate how "close" an individual may be to their FSS. The timeline of the experiment is shown in Fig 1. There are two groups of rats involved: a PPS group and a control group. The PPS group undergoes PPS and develops epilepsy before the end of the recording. The control group is not stimulated and they do not develop epilepsy before the end of the recording. Data from the control rats are used as a comparison to the PPS group. We demonstrate that our approach based on DNNs can successfully stage the EPG process and distinguish early from late EPG and that it generalizes to previously unseen rats.

2 RELATED WORK

2.1 Deep Learning for EEG analysis

Deep Learning (DL) techniques are commonly used in the analysis of EEG data in medical research. Example applications include the detection of Alzheimer's disease [4], autism [5], or Parkinson's disease [10]. In the context of epilepsy, DL has been applied for abnormal brain activity detection [22, 29] as well as seizure detection and prediction [7, 14, 28, 30, 33]. Roy *et al.* proposed a hybrid CNN and gated recurrent units (GRU) in classifying normal and abnormal brain activity, which takes time series EEG data as input and outputs the probability of being normal and abnormal, which is one of the first steps to understand the state of the brain activity in order to improve the accuracy of the diagnosis and the quality of patient care [22]. Tjepkema *et al.* explored different combinations of CNNs and recurrent neural networks (RNNs) as classifier to identify IEDs from scalp EEG [29]. Zhou *et al.* proposed a CNN-based approach to classify EEG time series data from different states, i.e., ictal, preictal, and interictal for the purpose of seizure detection [33]. They also compared the performance with time series and frequency-domain as input and found that frequency-domain input exhibits better potential for this task. Kiral-Kornek *et al.* proposed a DL-based approach for patient-specific seizure prediction by classifying intracranial EEG data in pre-ictal and interictal phases [14]. Thodoroff *et al.* proposed a neural network combining convolutional layers (conv-layers) with recurrent layers to detect seizure onset. Their network takes the image-based representation of EEG signals as input capturing spatial, spectral, and temporal features of patient-specific seizures [28]. Cho *et al.* compared the performance of different input modalities of EEG data with different DNN-based network architectures for seizure detection [7]. They concluded that the CNN with time-series EEG data, and the RNN with periodogram data resulted in the best performance. While these works have demonstrated the utility of DL for EEG analysis in the context of epilepsy, they have not addressed the challenging detection and staging of EPG prior to the FSS that we demonstrate here for the first time.

2.2 Interpretable DNNs

The interpretation of the reasoning of a neural network is crucial in medical applications, as it allows verification by human users and provides insights rather than just succumbing to a *black box*. Many studies have been done to address the interpretability of DNNs [13, 23, 25, 27, 31, 32]. Yosinski *et al.* developed a software tool for visualizing live feature extraction in the neural network by viewing the activation maps of different channels in different layers as well as by regularized optimization to generalize inputs that maximize the channel activation [31]. Simonyan *et al.* proposed to generate an input image that maximizes the output softmax probability of a given class. Meanwhile, a saliency map can be computed, which is the ranking of each pixel based on their contribution to the given class of a given sample [25]. Bach *et al.* proposed the Layer-wise Relevance Propagation (LRP), which understands the learning of the network by decomposing the output in terms of the input dimensions in a fashion that relates to Taylor decomposition [23]. Sturm *et al.* applied the LRP technique to visualize the frequency contribution to the classification result with EEG data [27]. Zhou *et al.* proposed the concept of class activation map (CAM), which can identify important regions in the inputs by propagating back the weights of the dense softmax layer to the inputs [32]. CAM is easy to deploy and provides more focused and localized discrimination. In this work, we also leverage CAM with 1-*d* EEG data to better visualize the network properties and the learned features.

2.3 EEG-based Biomarkers of Epileptogenesis

Over the last decades several studies have attempted to find EPG biomarkers in EEG signals. Li *et al.* and Bragin *et al.* focused on high-frequency oscillations (HFOs) in a rat epilepsy model with kainic acid (KA) injection [6, 17]. They found that the sooner HFOs appear after the injection, the higher the rate of spontaneous seizures in the chronic phase, and the shorter the latent period is, the more spontaneous seizures will occur. Milikovsky *et al.* focused on theta band activity and showed that a decreased theta power can be a robust feature in identifying EPG in five animal epilepsy models [19]. Andrade *et al.* investigated the role of sleep-wake disturbance

in EPG and found that there is a decrease of the dominant frequency and the duration of sleep spindles in a traumatic brain injury epilepsy model with generalized seizures [1]. Bentes *et al.* found that in stroke patients, the asymmetry in the background activity with the occurrence of IEDs are independent indicators of post-stroke epilepsy in the first year after stroke [3]. Sheybani *et al.* found that in a mouse model of epilepsy with kainate injection, the spatial propagation of a subgroup of spikes across the brain can be a reliable indicator of EPG as well as epilepsy in the chronic phase [24]. Lu *et al.* trained a DNN with the Fourier transformation of the time-series EEG data from a rat epilepsy model and showed that a decrease of power in theta band and an increase of power in frequencies over 100 Hz can be reliable indicators of EPG [18]. Rizzi *et al.* investigated the nonlinear dynamics of EEG signals and found a significant decrease of the so-called embedding dimension in early EPG that correlates with the severity of the ongoing EPG [21]. Here, we use an unbiased deep learning approach to study the EPG process to subdivide it into different stages and identify potential biomarkers to distinguish early and late phases of EPG.

3 METHODS

3.1 Animal Model

We use a mesial temporal lobe epilepsy with hippocampal sclerosis (mTLE-HS) rodent model, where epilepsy is electrically induced through PPS. Details have been described in [8]. Continuous single-channel EEG recordings from a depth electrode implanted in the dentate gyrus are collected from each rat from the beginning of the implantation until the FSS, which indicates the manifestation of epilepsy. The 24/7 recordings enable us to continually monitor the entire EPG prior to the FSS. There are two groups of rats involved in this study, 1) seven rats had PPS and developed epilepsy before the end of recording, which we denote as PPS rats, 2) three rats did not get PPS stimulation and did not develop epilepsy by the end of recording, which we denote as control rats. In the PPS group, the average EPG phase is 4 weeks (range 1 – 7 weeks). The EPG phase is terminated by the FSS. The timelines for both group are shown in Fig. 1. Training data are taken from the three highlighted periods from PPS rats for the three-class classification task. We define the three classes to be the Baseline class (*BL*) – green, the *early EPG* class – blue, and the *late EPG* class – orange. The data from the control rats are used only for testing the model trained on the PPS group. The total available number of recordings from each rat is summarized in Table 1 and Table 2.

3.2 EEG Data Preprocessing

The data acquisition was achieved through wireless EEG transmitters with a sampling rate of 512 Hz and a band-pass filter between 0.5 - 160 Hz as well as a notch filter at 50 Hz. Occasionally, EEG artifacts can appear as extreme amplitude values and signal loss due to electronic interference and weak transmission. To combat this problem, we first applied a MATLAB function, i.e., `filloutliers` [1] with the parameters `method = 'pchip'; movmethod = 'movmedian'; window = 50` to filter out unrealistic extreme values. Then, the continuous recordings are divided into five-second long non-overlapping segments. To manage data loss, we discarded any five-second segments with more than 20 % data loss. As a result, we discarded around 5% of the total recordings. The remaining segments were eligible for the DNN training.

[1] https://www.mathworks.com/help/matlab/ref/filloutliers.html

Table 1: Summary of the data collections from PPS rats in hours (hrs).

rat ID	PPS 1	PPS 2	PPS 3	PPS 4	PPS 5	PPS 6	PPS 7
BL (hrs)	162	160	149	82	163	164	157
EPG (hrs)	700	508	400	140	1568	173	648

Table 2: Summary of the data collections from control rats in hours (hrs).

rat ID	Ctr 1	Ctr 2	Ctr 3
in total (hrs)	1536	2140	2248

3.3 DNN Architecture

We use a deep residual neural network with 16 blocks with residual connections (res-block), as shown in Fig. 2, inspired by [11]. The model takes five-second long EEG segments as input and outputs the probability over three classes, i.e., BL, early EPG, and late EPG. We keep the design of each res-block as in [11], where each res-block consists of two conv-layers, batch-normalization, dropout, and ReLU non-linear activation. The number of channels in the first conv-layer and the first block is 16, and it increases by a factor of 2 in every four blocks. There are two branches in each block: one goes through convolution, batch-normalization, ReLU activation and dropout; the other, called skip connection, simply goes through max-pooling. They are combined in an additive manner at the end of the block before passing through the batch-normalization and ReLU activation. To reduce the dimensionality of the feature maps, we use a stride of two in the second conv-layer and the max-pooling layer in every other block starting from the second block. The output of the last conv-layer is fed to the global average pooling (GAP) operation, which is followed by a dense layer with three output units with softmax non-linear activation. The dropout rate is 0.2 everywhere in the graph.

3.4 Class Activation Map

Proposed by Zhou *et al.*, the class activation map is a method to visualize the "importance" of different regions of the input for the classification decision. It takes advantage of the global average pooling (GAP) after the last conv-layer, and assigns different weights to each squashed feature map. To be specific, the k-th feature map from the last conv-layer, denoted as f_k, which has shape [$height$, $width$]. The GAP layer takes the mean activation of each f_k, and the resulting k-th feature map F_k is $\frac{1}{N}\sum_{i,j} f_k(i,j)$, where N is the total number of elements of f_k. It reduces the dimension by the factor of $height \times width$. Then, for a given class c, the input to the softmax layer, S_c, is a weighted linear combination of all the feature

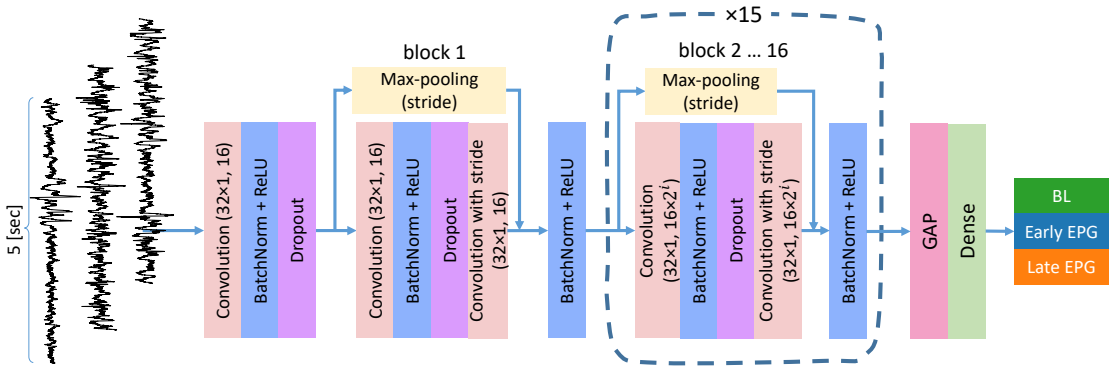


Figure 2: The DNN structure used in this study. The network takes a mini-batch of five-second segments as input and outputs the probability over the three classes. GAP: global average pooling. BL: Baseline

maps, which is computed by

$$S_c = \sum_k w_k^c \frac{1}{N} \sum_{i,j} f_k(i,j) \propto \sum_{i,j} \sum_k w_k^c f_k(i,j), \tag{1}$$

where w_k^c denotes the *importance* of f_k for class c. Finally, the softmax probability for class c can be computed as $\frac{\exp S_c}{\sum_c \exp S_c}$. Then, when the training is finished, the class activation map for class c at position (i, j), $\mathrm{CAM}_c(i, j)$, is given by

$$\mathrm{CAM}_c(i,j) = \sum_k w_k^c f_k(i,j). \tag{2}$$

Hence, $S_c = \sum_{i,j} w_k^c \mathrm{CAM}_c(i,j)$, and the weights w_c are fixed after the training. Then, $\mathrm{CAM}_c(i, j)$ indicates the importance of the activation at the position (i, j) contributing to the class c.

3.5 DNN Training and Evaluation

We apply a seven-fold leave-one-out cross-validation (LOO-CV) scheme, where the network is trained with the data from six out of seven rats in the PPS group. Specifically, in each fold, we withhold the data from one rat as the test set, and the data from other six rats form the training and the validation sets with a train-validation-split of 8:2. This procedure is repeated seven times, and each time we hold out a different rat for testing. This is highly relevant to test the generalization ability of the classifier to unseen data from unseen subjects. We randomly select 25 hours from a three-day window from each phase for training and validation, shown as the shaded periods in Fig. 1. The choice of 25 hours is a reasonable trade-off between computational cost and performance from empirical experience, since we also experiment using all data from the three day periods and it increases the total training time by a factor of three and no significant improvement regarding the classification performance is found. Our DNN model is implemented in Tensorflow and trained with an NVIDIA GeForce RTX 2080 Ti GPU and one epoch of training takes 35 minutes on average. After the network is trained, we test it with all the data from those three-day periods (shown in Fig. 1) of the previously withheld rat. We report results as the average across all seven LOO test trials.

To evaluate the performance, we compute the receiver operating characteristic (ROC) curve in the multi-class scenario, where the ROC curve is computed for each class in a one-vs-all manner. The area under the ROC curve is a scalar value indicating the goodness of the trained classifier. Several other performance metrics including precision, recall, and F1-score are also computed. These metrics are given by:

$$\begin{aligned}
\text{precision} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\
\text{recall} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\
\text{F1-score} &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \\
\text{accuracy} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}},
\end{aligned}$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative numbers, respectively. We also compare our results with several baseline network structures: a feed-forward neural network (FNN), a deep convolutional neural network (DCNN) [26], EEGNet [16], and one variant of our proposed model with only four blocks, which we denote as Proposed-4block.

The FNN used in this work is a straight forward multi-layer perceptron with three dense layers equipped with 1024, 256, and 128 units per layer. Each dense layer is regularized with L_2 penalty with a factor of 0.01 and followed by a batch-normalization layer and a dropout (rate=0.5) layer. The nonlinear activation is ReLU in this model.

Sors *et al.* proposed the DCNN for sleep staging with single-channel EEG. Compared to the original architecture, we made several changes to adapt to the training data format we have in our experiment. First, due to our input being shorter (five-second segments under 512 Hz sampling rate, which yields 2560 data points per sample) than theirs (15 000 data points), we reduced the number of conv-layer from twelve to nine: five (instead of six) conv-layers with 128 output channels and four (instead of six) conv-layers with 256 output channels. Each conv-layer has stride 2 to sub-sample the feature map. The architecture is conv (7×1, 128, stride 2) – conv (7×1, 128, stride 2) – conv (7×1, 128, stride 2) – conv (7×1, 128, stride 2) – conv (7×1, 128, stride 2) – conv (5×1, 256, stride 2) – conv (5×1, 256, stride 2) – conv (5×1, 256, stride 2) – conv (3×1, 256,

stride 2) – flatten – fully-connected (units=100) – fully-connected (units=3). We kept other training parameters identical to the original paper.

Lawhern *et al.* proposed the original EEGNet for EEG classification in multiple brain-computer interfaces. The EEG snippets used in their evaluation are multi-channel event related potential (ERPs) recorded from surface EEG setups, band-pass filtered between 1-40 Hz, downsampled to 128 Hz, and focused on 1 to 2 seconds around the event onset. The original EEGNet demonstrates good generalization to EEG classification among different experiment diagrams even though the total number of parameters is two orders of magnitude smaller than the baseline methods evaluated in their work. To adapt EEGNet to our task, we made several changes to the architecture while keeping layers such as batch-normalization, dropout, exponential linear unit (ELU) activation function, and average pooling unchanged: 1) We expanded the width of the convolutional filter from 64 to 256, which is half of our sampling rate as suggested in the original paper. 2) We used three instead of two layers of convolution while omitting the depth-wise convolution, since our data is single-channel. Unfortunately, the classification accuracy of this modified EEGNet (henceforth denoted EEGNet1) does not exceed chance-level. One contributing factor might be the low number of trainable parameters. In total, EEGNet1 only has 223 323 learnable parameters, which is considerably fewer than our proposed model. To make the total number of trainable comparable to ours, we increased the number of conv-layers and the number of filters in each layer. This is essentially equivalent to a relatively shallow CNN (7 conv-layers compared to 33 layers in our proposed model) with very wide convolutional filters, which we denote as EEGNet2. The resulting structure of EEGNet2 is conv $(256 \times 1, 16)$ – batch-normalization – conv $(256 \times 1, 16)$ – batch-normalization + ELU + average-pooling + dropout – conv $(256 \times 1, 32)$ – batch-normalization + ELU + average-pooling + dropout – conv $(256 \times 1, 32)$ – batch-normalization + ELU + average-pooling + dropout – conv $(256 \times 1, 64)$ – batch-normalization + ELU + average-pooling + dropout – conv $(256 \times 1, 64)$ – batch-normalization + ELU + average-pooling + dropout – flatten – fully-connected (units=3). As a result, the EEGNet2 has a total number of 4 195 107 parameters, which is comparable to that of our proposed model (4 200 048). However, the results show that with the same amount of training data and training time, both versions of EEGNets, i.e., EEGNet1 and EEGNet2 perform at chance-level. Thus, their performance measures were omitted in the performance report.

4 EXPERIMENTS AND RESULTS

Table 3 shows the performance summary of our proposed model in comparison to the baseline methods. The reported performance metrics are averaged for each class as well as a macro-average of all classes across all LOO test trials. Our proposed method obtains the best performance in almost all evaluated metrics compared to the baseline methods. Notably our proposed-4block model still obtains better performance than FNN and DCNN, even though the number of trainable parameters is more than 20 times smaller. Compared to the full-size proposed model, the Proposed-4block model suffers from a slight performance degradation. From the class-wise performance, we can see that, in general, the BL class is easier for the networks to classify as shown by the highest average performance among the three classes in all models.

4.1 Prediction Aggregation and ROC Analysis

To gather statistics of the estimated class membership over a longer time period, we apply a prediction aggregation technique as proposed in our previous study [18]. Essentially, we apply a linear average aggregation of the resulting softmax probability across multiple consecutive five second data segments such that the probabilities of each class are accumulated across a longer period of time. Figure 3 shows the averaged AUCs of the three classes across all LOO test trials **with** and **without** the prediction aggregation (Fig. 3A and Fig. 3B) as well as the effect of the pooling length used in the prediction aggregation (Fig. 3C). In general, the network can distinguish BL segments better than the other two classes as shown by the highest average AUC under the ROC curve among the three classes, with or without the prediction aggregation. Prediction results for the control group are only marginally better than chance, suggesting that the network really detects changes in brain activity patterns due to the PPS, rather than any changes triggered by the initial electrode implantation that are independent of the PPS. Prediction aggregation over one hour increases the average AUC of the baseline, early, and late EPG classes by 0.1, 0.12, and 0.11, respectively.

To study the benefits of aggregation in more detail, we compute the AUCs for various aggregation lengths in each LOO test trial, i.e., 5 seconds, 30 seconds, one, two, five, ten, 20, 30, and 60 minutes. The average AUC as a function of the aggregation lengths is depicted in Fig 3C. It reflects the inter-rat variability in the three-class classification with our proposed network, i.e., the AUC starts at different levels of confidence without prediction aggregation (the first data points from all rats). The figure shows that with an increasing pooling length, the average AUC increases in all LOO test trials. To be specific, with one hour of aggregation, the average AUC improved by 0.12 (a maximum of 0.18 and a minimum of 0.06). Hence, aggregating the softmax output from the network across multiple consecutive segments captures trends across a longer period, which is essential for distinguishing different classes in our task. Aggregation over even longer time periods (>1 hour) might be able to further improve performance.

4.2 Disease Progression

EPG is a gradual process, but above we treated EPG detection and staging as a discrete classification problem by defining (somewhat arbitrarily) the first three days after the stimulation as the early EPG phase, and the last three days before the FSS as the late EPG phase. The data from the period in between these two phases has not been considered so far. In the following, we analyze samples from this intermediate period and study how the network, which has been trained to distinguish Baseline, early and late EPG phases, will classify them. Specifically, we consider the estimated probability for each class, denoted as the *class score*, throughout the whole recording period from a randomly picked pre-trained model from the LOO cross-validation scheme, which we call "Pretrained-1" model. Here, we are interested in the general trend rather than the

Table 3: Performance across all leave-one-out test trials with one hour of prediction aggregation. Evaluation metrics are reported in class-wise average and overall average for each model. Numbers are shown in *mean* ± *std*.

Model	Class	Precision	Recall	F1-score	Accuracy	# trainables
FNN	0	0.51 ± 0.08	0.67 ± 0.13	0.57 ± 0.06	0.44 ± 0.06	2 920 963
	1	0.49 ± 0.07	0.65 ± 0.07	0.55 ± 0.02	0.43 ± 0.04	
	2	0.43 ± 0.06	0.63 ± 0.16	0.49 ± 0.04	0.39 ± 0.03	
	average	0.47 ± 0.03	0.65 ± 0.04	0.53 ± 0.03	0.42 ± 0.03	
DCNN [26]	0	0.47 ± 0.13	0.54 ± 0.25	0.46 ± 0.18	0.46 ± 0.13	1 607 187
	1	0.43 ± 0.32	0.40 ± 0.30	0.41 ± 0.31	0.41 ± 0.08	
	2	0.35 ± 0.23	0.35 ± 0.28	0.33 ± 0.23	0.40 ± 0.07	
	average	0.42 ± 0.11	0.43 ± 0.22	0.40 ± 0.18	0.42 ± 0.07	
Proposed-4block	0	0.70 ± 0.14	0.88 ± 0.04	0.78 ± 0.10	0.66 ± 0.14	82 912
	1	0.43 ± 0.06	0.68 ± 0.18	0.53 ± 0.10	0.41 ± 0.05	
	2	0.51 ± 0.05	**0.82 ± 0.13**	0.62 ± 0.01	0.47 ± 0.02	
	average	0.55 ± 0.04	0.79 ± 0.09	0.64 ± 0.05	0.51 ± 0.05	
Proposed model	0	**0.85 ± 0.17**	**0.96 ± 0.02**	**0.90 ± 0.10**	**0.84 ± 0.17**	4 200 048
	1	**0.69 ± 0.12**	**0.81 ± 0.17**	**0.74 ± 0.14**	**0.64 ± 0.15**	
	2	**0.71 ± 0.33**	**0.74 ± 0.32**	**0.72 ± 0.31**	**0.71 ± 0.22**	
	average	**0.75 ± 0.15**	**0.84 ± 0.12**	**0.78 ± 0.14**	**0.73 ± 0.14**	

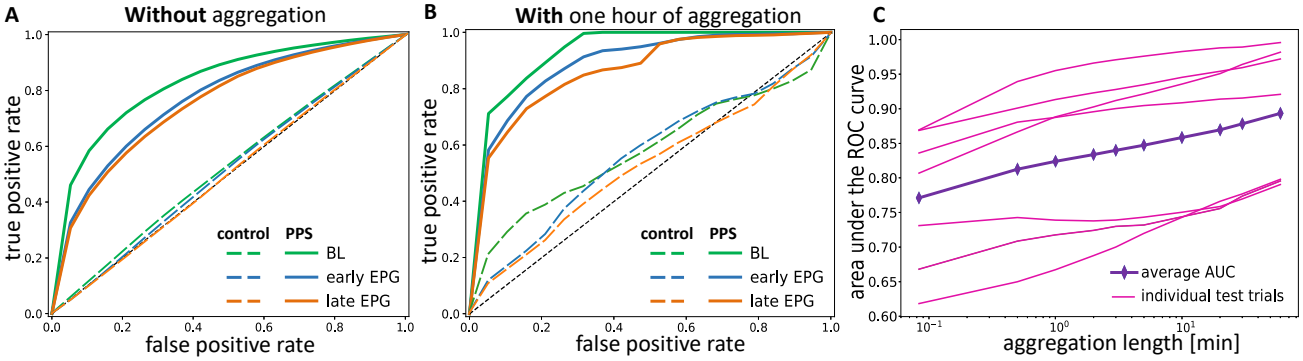


Figure 3: Network performance across all test trials within the PPS and the control group. A. Average ROC curves of multiple classes without aggregation within the PPS group and the control group. The AUC for the three classes of PPS rats are 0.83, 0.77 and 0.75 (solid lines) and those of the control rats are 0.52, 0.51, and 0.50 (dashed lines). B. Average ROC curves of multiple classes with aggregation over one continuous hour within the PPS and the control group. The AUC of the three classes for the PPS rats (solid lines) are 0.93, 0.89, and 0.86, and those of the control rats are 0.58, 0.56, and 0.53 (dashed lines). C. The AUC as a function of the aggregation length in all individual PPS LOO test trials (magenta lines) and the average AUC of all classes across all trials (purple with diamonds). ROC: receiver operating characteristic. AUC: area under the curve.

classification accuracy, so the training data were also included. The progression of class scores from two example **PPS** rats and one **control** rat are shown in Fig. 4. One of the PPS rats (PPS 1) has a relatively long EPG duration (30 days) and the other (PPS 4) has a short EPG duration (6 days). The control rat (Ctr 1) has 64 days of recordings in total.

Several findings are evident in the data for the PPS rats in Fig. 4A,B. First, the Baseline score is high during the entire baseline period and drops to small values during the EPG phase. Second, with the beginning of the EPG phase, the early EPG score increases and then gradually decreases towards the late EPG phase. Third, conversely, the late EPG score is low during baseline and the beginning of EPG and then gradually increases towards the late EPG phase. Fourth, in some animals we observe a circadian rhythm in the early and late EPG scores during the transition period between early and late EPG (compare Fig. 4A). These findings are in sharp contrast to those for the control rats. In their case, the late EPG score remains low throughout the entire recording period, in line with these animals not developing epilepsy during the experiment (compare Fig. 4C).

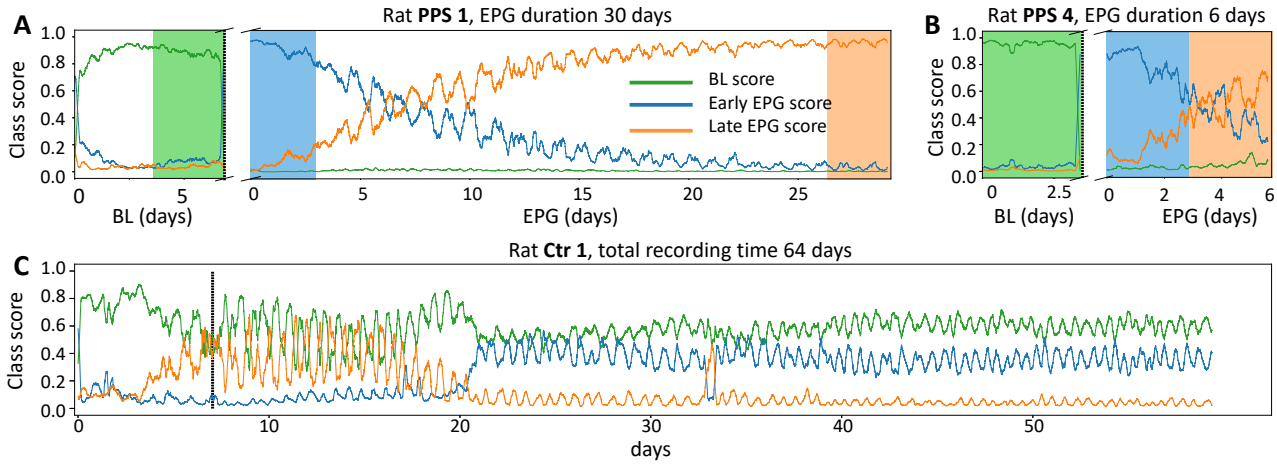


Figure 4: Class scores from two PPS rats (A,B) and one control rat (C) during the entire recording. The vertical black dashed lines indicate the time when the PPS rats started receiving PPS, while control rats did not.

4.3 Feature Representation

The interpretation of EEG signals is always challenging, since they are highly variable — especially across subjects. Analyzing and understanding the discriminative features learned by a DNN model can give valuable insights as to what distinguishes the classes. This can be particularly helpful in medical applications, where the differences between classes many not be easily spotted — even by the expert eye. Here, we present the feature representations learned by the network. Using the Pretrained-1 model, we passed all the data from all seven rats through the network and computed the average activation of the last conv-layer for each class, shown on the top left of Fig. 5. We can see that there is a group of feature channels that are very active. Most importantly some of these feature channels are most active for one class but not the others and some extract features that contribute to more than one class. Next, we identified several channels that were highly active for each class and plotted the EEG segments that maximally activate them. Interestingly, we found several feature channels responding to very distinctive features such as spikes in channel 3, spike-and-slow-waves in channel 9, spindles and HFOs in channel 15, theta rhythm in channel 16, delta wave plus low beta in channel 21, etc. From this we can conclude that before the softmax layer, the network has already extracted class-specific features that are clinically meaningful.

To further elucidate which parts of the input contribute most to the classification of the different EPG stages, we leverage the CAM visualization while manipulating the assigned labels for the EEG segments. Taking Pretrained-1 model, we freeze the weights and for a given sample, we assign in turn the three different labels. Then, by computing the CAM of the given sample under the assigned label, we trace back which parts of the given five second input segment most support (> 80-th percentile) the assigned classification. The results are shown in Fig. 6. Indeed, the CAMs for the sample vary depending on the given label. There are several interesting features that the network has discovered. First, the BL class is most supported by low-amplitude waves, and many downwards deflections. Second, sharp waves contribute to both EPG classes, but the difference lies in the width of the wave forms. While an early EPG classification is supported by narrow spikes, or spike-like waves, a late EPG classification is supported by somewhat wider sharp waves.

5 CONCLUSION

We have proposed a DNN model for single-channel intracranial EEG classification to better understand the progression of epileptogenesis (EPG). Specifically, our aim was to stage the EPG process prior to the first spontaneous seizure (FSS), which could facilitate early intervention **before** an epilepsy becomes manifest. In previous work, we had already shown that a DNN can learn to distinguish EEG data from before and after the epilepsy-inducing stimulation with high discrimination and generalization ability [18]. Here, we have sought to answer a) whether we can further distinguish different stages of EPG before the FSS, and b) what EEG features would be representative for each stage. To this end, we have trained a DNN model with five-second EEG segments recorded from three phases in a rodent epilepsy model [8]: three days before the PPS (Baseline, BL), three days shortly after the PPS (early EPG), and three days immediately before the FSS (late EPG). We have evaluated our approach in a LOO scheme to test the generalization ability of the model to data from unseen rats. To pool evidence over larger time windows, we applied a prediction aggregation method as in previous work [18]. We also compared the performance of our model to four other models, specifically an FNN model, a DCNN model [26], the well-known EEGNet [16], and a reduced version of our model with 50 times fewer parameters. In an extensive performance evaluation, we showed that our proposed model yielded the best results and could distinguish different EPG stages with high accuracy. Furthermore, we showed that the network learns to extract meaningful EEG features to perform the classification.

Various challenges will need to be overcome, in order to translate our findings to human patients. First, the rodent model we have

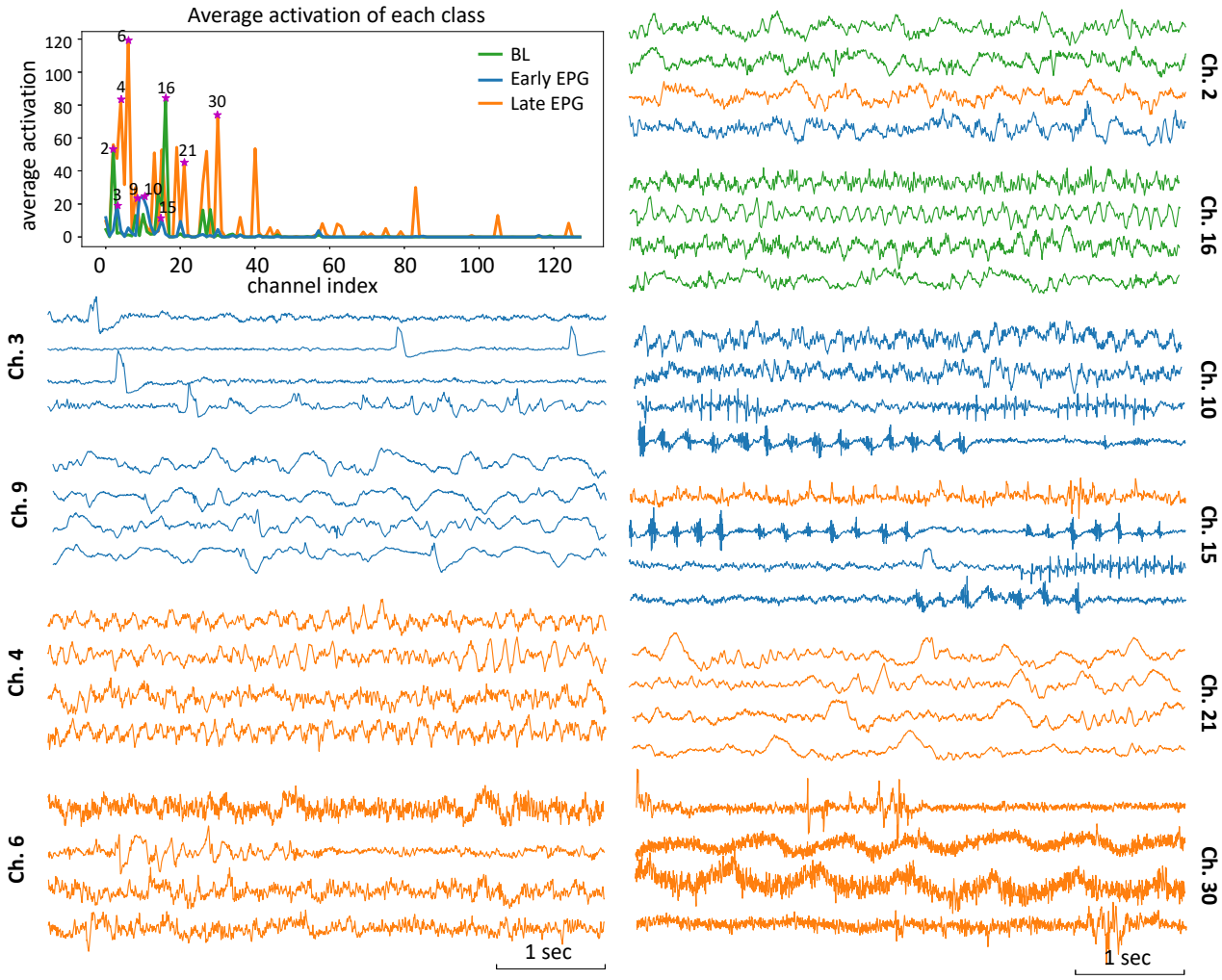


Figure 5: Normalized average activation of the last conv-layer by class (top left). Examples of five second EEG samples that maximize the activation of certain channels in the last conv-layer. Color indicates a sample's class label. Scale bar represents 1 second.

used provides quasi ideal conditions, supplying high quality, 24/7 intracranial recordings directly from the affected brain region. It is unclear whether similar results could be achieved with surface EEG recordings from a diverse set of human patients. The second challenge is that epilepsy is typically diagnosed only *after* the FSS. In order to attempt early detection of EPG as we have demonstrated here in human patients, one would have to obtain recordings from patients *before* the FSS. This requires monitoring a population of patients with a sufficiently high risk of developing epilepsy, which is challenging. Third, our approach relies on a large data set comprising around-the-clock recordings over several weeks for each individual. Acquiring similar data from a (homogeneous) patient population would be very difficult. It is an open question, how much data would be required to allow accurate classification and good generalization. Fortunately, in our experiments, pooling data over one hour already provided very good results. Such a time span appears manageable in clinical practice. Finally, even if EPG could be detected and staged reliably in human patients at risk of developing epilepsy, it is far from clear which forms of early intervention would be effective in modifying or halting the disease development. In fact, such interventions will likely have to depend on the specific type of epilepsy and be adapted to individual patients. In the future, machine learning may also support physicians in this challenging task.

ACKNOWLEDGMENTS

This work is supported by the China Scholarship Council (CSC, No. [2016]3100), the LOEWE Center for Personalized Translational Epilepsy Research (CePTER), and the Johanna Quandt Foundation. The authors would like to thank Markus Ernst for proofreading the paper and providing valuable feedback.

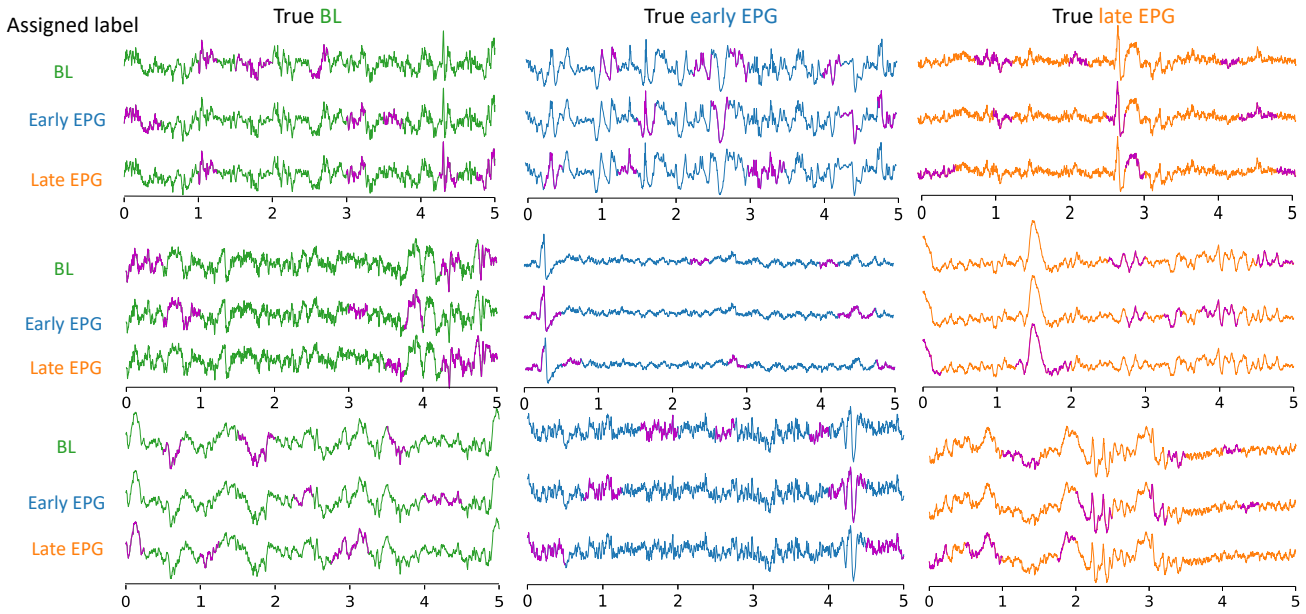


Figure 6: Identifying informative regions in the five-second long input via CAM. The main color of a trace corresponds to its true label. The areas highlighted in magenta most strongly support the assigned classification (>80-th percentile).

REFERENCES

[1] Pedro Andrade, Jari Nissinen, and Asla Pitkänen. 2017. Generalized seizures after experimental traumatic brain injury occur at the transition from slow-wave to rapid eye movement sleep. *Journal of neurotrauma* 34, 7 (2017), 1482–1487.

[2] AJ Becker. 2018. Animal models of acquired epilepsy: insights into mechanisms of human epileptogenesis. *Neuropathology and applied neurobiology* 44, 1 (2018), 112–129.

[3] Carla Bentes, Hugo Martins, Ana Rita Peralta, Carlos Morgado, Carlos Casimiro, Ana Catarina Franco, Ana Catarina Fonseca, Ruth Geraldes, Patrícia Canhão, Teresa Pinho e Melo, et al. 2018. Early EEG predicts poststroke epilepsy. *Epilepsia open* 3, 2 (2018), 203–212.

[4] Xiaojun Bi and Haibo Wang. 2019. Early Alzheimer's disease diagnosis based on EEG spectral images using deep learning. *Neural Networks* 114 (2019), 119–135.

[5] William J Bosl, Helen Tager-Flusberg, and Charles A Nelson. 2018. EEG analytics for early detection of autism spectrum disorder: a data-driven approach. *Scientific reports* 8, 1 (2018), 1–20.

[6] Anatol Bragin, Charles L Wilson, Joyel Almajano, Istvan Mody, and Jerome Engel Jr. 2004. High-frequency oscillations after status epilepticus: epileptogenesis and seizure genesis. *Epilepsia* 45, 9 (2004), 1017–1023.

[7] Kyung-Ok Cho and Hyun-Jong Jang. 2020. Comparison of different input modalities and network structures for deep learning-based seizure detection. *Scientific Reports* 10, 1 (2020), 1–11.

[8] Lara S Costard, Valentin Neubert, Morten T Venø, Junyi Su, Jørgen Kjems, Niamh MC Connolly, Jochen HM Prehn, Gerhard Schratt, David C Henshall, Felix Rosenow, et al. 2019. Electrical stimulation of the ventral hippocampal commissure delays experimental epilepsy and is associated with altered microRNA expression. *Brain Stimulation* 12, 6 (2019), 1390–1401.

[9] Jerome Engel Jr and Asla Pitkänen. 2020. Biomarkers for epileptogenesis and its treatment. *Neuropharmacology* 167 (2020), 107735.

[10] Faraz Faghri, Sayed Hadi Hashemi, Hampton Leonard, Sonja W Scholz, Roy H Campbell, Mike A Nalls, and Andrew B Singleton. 2018. Predicting onset, progression, and clinical subtypes of Parkinson disease using machine learning. *bioRxiv* (2018), 338913.

[11] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine* 25, 1 (2019), 65.

[12] Mark R Keezer, Sanjay M Sisodiya, and Josemir W Sander. 2016. Comorbidities of epilepsy: current concepts and future perspectives. *The Lancet Neurology* 15, 1 (2016), 106–115.

[13] Pieter Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. 2018. Learning how to explain neural networks: PatternNet and PatternAttribution. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings* (2018), 1–12.

[14] Isabell Kiral-Kornek, Subhrajit Roy, Ewan Nurse, Benjamin Mashford, Philippa Karoly, Thomas Carroll, Daniel Payne, Susmita Saha, Steven Baldassano, Terence O'Brien, et al. 2018. Epileptic seizure prediction using big data and deep learning: toward a mobile system. *EBioMedicine* 27 (2018), 103–111.

[15] Patrick Kwan and Martin J Brodie. 2000. Early identification of refractory epilepsy. *New England Journal of Medicine* 342, 5 (2000), 314–319.

[16] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering* 15, 5 (2018), 056013.

[17] Lin Li, Mayur Patel, Joyel Almajano, Jerome Engel Jr, and Anatol Bragin. 2018. Extrahippocampal high-frequency oscillations during epileptogenesis. *Epilepsia* 59, 4 (2018), e51–e55.

[18] Diyuan Lu, Sebastian Bauer, Valentin Neubert, Lara Sophie Costard, Felix Rosenow, and Jochen Triesch. 2019. A Deep Residual Neural Network Based Framework for Epileptogenesis Detection in a Rodent Model with Single-Channel EEG Recordings. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 1–6.

[19] Dan Z Milikovsky, Itai Weissberg, Lyn Kamintsky, Kristina Lippmann, Osnat Schefenbauer, Federica Frigerio, Massimo Rizzi, Liron Sheintuch, Daniel Zelig, Jonathan Ofer, et al. 2017. Electrocorticographic dynamics as a novel biomarker in five models of epileptogenesis. *Journal of Neuroscience* 37, 17 (2017), 4450–4461.

[20] Asla Pitkänen and Jerome Engel. 2014. Past and present definitions of epileptogenesis and its biomarkers. *Neurotherapeutics* 11, 2 (2014), 231–241.

[21] Massimo Rizzi, Claudia Brandt, Itai Weissberg, Dan Z Milikovsky, Alberto Pauletti, Gaetano Terrone, Alessia Salamone, Federica Frigerio, Wolfgang Löscher, Alon Friedman, et al. 2019. Changes of dimension of EEG/ECoG nonlinear dynamics predict epileptogenesis and therapy outcomes. *Neurobiology of disease* 124 (2019), 373–378.

[22] Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. 2019. ChronoNet: a deep recurrent neural network for abnormal EEG identification. In *Conference on Artificial Intelligence in Medicine in Europe*. Springer, 47–56.

[23] Bach Sebastian, Binder Alexander, Montavon Grégoire, Klauschen Frederick, Müller Klaus-Robert, Samek Wojciech, and Suarez Oscar Deniz. [n.d.]. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *Plos One* 10, 7 ([n. d.]), e0130140.

[24] Laurent Sheybani, Gwenaël Birot, Alessandro Contestabile, Margitta Seeck, Jozsef Zoltan Kiss, Karl Schaller, Christoph M Michel, and Charles Quairiaux. 2018. Electrophysiological evidence for the development of a self-sustained large-scale epileptic network in the kainate mouse model of temporal lobe epilepsy. *Journal of Neuroscience* 38, 15 (2018), 3776–3791.

[25] K Simonyan, A Vedaldi, and A Zisserman. 2017. Deep inside convolutional networks: visualising image classification models and saliency maps. CoRR 2013; abs/1312.6034. *arXiv preprint arXiv:1312.6034* (2017).

[26] Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean-François Payen. 2018. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control* 42 (2018), 107–114.

[27] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2016. Interpretable deep neural networks for single-trial EEG classification. *Journal of neuroscience methods* 274 (2016), 141–145.

[28] Pierre Thodoroff, Joelle Pineau, and Andrew Lim. 2016. Learning robust features using deep learning for automatic seizure detection. In *Machine learning for healthcare conference*. 178–190.

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

[30] KD Tzimourta, AT Tzallas, N Giannakeas, LG Astrakas, DG Tsalikakis, and MG Tsipouras. 2018. Epileptic seizures classification based on long-term EEG signal wavelet analysis. In *Precision medicine powered by pHealth and connected health*. Springer, 165–169.

[31] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding Neural Networks Through Deep Visualization. *Computer Science* (2015).

[32] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

[33] Mengni Zhou, Cheng Tian, Rui Cao, Bin Wang, Yan Niu, Ting Hu, Hao Guo, and Jie Xiang. 2018. Epileptic seizure detection based on EEG signals and CNN. *Frontiers in neuroinformatics* 12, 95 (2018).

HUMAN-EXPERT-LEVEL BRAIN TUMOR DETECTION USING DEEP LEARNING WITH DATA DISTILLATION AND AUGMENTATION

Diyuan Lu[⋆] *Nenad Polomac*[†] *Iskra Gacheva*[†] *Elke Hattingen*[†] *Jochen Triesch*[⋆]

[⋆] Frankfurt Institute for Advanced Studies, Frankfurt am Main, Germany
[†] Institute for Neuroradiology at Frankfurt University Hospital, Frankfurt am Main, Germany

ABSTRACT

The application of Deep Learning (DL) for medical diagnosis is often hampered by two problems. First, the amount of training data may be scarce, as it is limited by the number of patients who have acquired the condition. Second, the training data may be corrupted by various types of noise. Here, we study the problem of brain tumor detection from magnetic resonance spectroscopy (MRS) data, where both types of problems are prominent. To overcome these challenges, we propose a new method for training a deep neural network that distills particularly representative training examples and augments the training data by mixing these samples from one class with those from the same and other classes to create additional training samples. We demonstrate that this technique substantially improves performance, allowing our method to achieve human-expert-level accuracy with just a few thousand training examples.

***Index Terms*—** brain tumor, magnetic resonance spectroscopy (MRS), noisy labels, deep neural network, data augmentation

1. INTRODUCTION

Modern machine learning (ML) approaches based on deep neural networks (DNNs) have recently obtained impressive results in a range of classification tasks, sometimes even outperforming human experts. These successes are based on, amongst others, 1) better learning algorithms, 2) fast computing hardware, and 3) large, carefully annotated data sets. This has motivated a range of applications in the healthcare domain such as cardiovascular disease classification [1], tumor detection, tumor segmentation, tumor progression estimation [2–6], tumor grade classification [6], etc. However, acquiring the required large labeled data sets is often hard to achieve or expensive in certain medical applications where the numbers of patients may be quite small. Typical data sets often contain only hundreds or thousands of samples, while modern ML approaches often require the estimation of many millions of free parameters. Fitting a model with many free parameters to a small set of training samples will likely lead to over-fitting and poor generalization of the learned model. This problem is aggravated if the training data are corrupted by different kinds of noise, which is often unavoidable in biomedical data.

Here, we study the problem of brain tumor detection from magnetic resonance spectroscopy (MRS) data. In clinical practice, MRS is a common tool to identify a brain tumor and distinguish it from other medical conditions. MRS measures the resonant frequency shift of a chemically bound hydrogen atom (i.e., a proton), which characterizes different physiological or pathological brain metabolites. There has been increasing interest in MRS for clinical use because of the semiautomatic data acquisition, processing and quantification [6, 7]. While the interpretation of spectra is traditionally based on the size and location of certain peaks, we here use a novel approach by analysing the pattern of the MR spectrum as a whole in an unbiased fashion with DNNs.

A common problem with MRS data is noise. Noise sources include head movement during the procedure, or baseline distortions of the spectrum. Additionally, labels are only provided per patient and not per voxel, which could introduce labeling noise as spectra from the tumor-affected hemisphere can be falsely labeled as "tumor" even though they contain healthy brain tissue. Given the ubiquity and importance of coping with noisy labeling, many works on this topic have been published [8–11]. Starting learning from a small set of expert validated labels is one promising direction [8, 10]. Another direction is to design models that learn directly with noisy labels [10, 11]. For example, [10] uses a co-teaching framework where two DNNs were trained simultaneously with noisy labeling, or [11] discards samples that contribute negatively to the training performance.

Scarcity of training data can be another big hurdle when applying DL methods to medical problems. Data augmentation is a common approach to alleviate this. The new samples can be generated by training a generative model as in [12, 13], or by blending two or more of the original training samples as in [14, 15]. Here, we propose a novel dual-step framework including: 1) a data distillation step, which determines representative training examples, and 2) classifier training with data augmentation. In a nutshell, our method works by automatically identifying data points that are "easy" to classify through a distillation network. Then, these data samples form

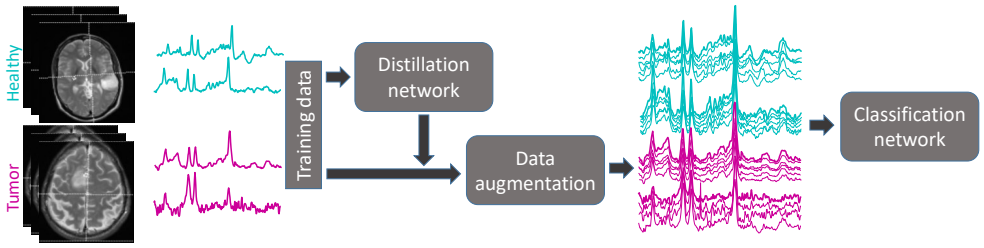


Fig. 1. Overview of the proposed approach. 1. A data distillation network automatically selects representative samples. 2. These distilled samples form the basis of the data augmentation process synthesizing the training samples for the a final classifier.

the basis of a data augmentation process that generates more training samples for training the final classifier. Specifically, new training samples are generated by mixing the distilled "easy" samples with randomly selected data points from the same or the other class. The proposed framework is illustrated in Fig. 1. Notably, it does not require human supervision to carefully annotate a small data set as prototype samples, but rather learns directly on the noisy labeled data as in [10, 11]. Note that in this work, the distillation network and the final classifier have the same structure, but this need not be the case. Overall, we make the following contributions: 1) We apply deep learning to MRS data from a cohort of patients with multiple medical conditions. 2) We propose a framework for tumor classification based on MRS data that combines DNNs with a novel data distillation and augmentation procedure to combat scarcity of the training data and label noise. 3) We quantify the classification performance of human expert neuroradiologists on the same MRS data set and demonstrate that our approach achieves human-expert-level performance.

2. DATASET

1H-MR-spectroscopy data from 435 patients recorded in the Institute for Neuroradiology of the University Hospital, Frankfurt during the time interval from 01/2009 to 3/2019 were reviewed retrospectively. These patients were suffering from either glial or glioneuronal first diagnosed tumors (the *tumor* group) or other non-neoplastic lesions, e.g., demyelination, gliosis, focal cortical dysplasia, enlarged Virchow-Robin spaces or similar (the non-tumor/*healthy* group). The tumor group included all spectra from the tumor-affected hemisphere. The non-tumor group consisted of all spectra from both hemispheres of the patients. As a result, 7442 spectra (3388 non-tumor and 4054 tumor) were selected for further analysis. The obtained MR spectra are represented as column vectors (288×1) and reflect the chemical shift positions in ppm indicating various metabolites.

3. METHODS

Data distillation. To automatically distill the "easy" samples from the data set with noisy labels, we propose a data distillation setup, which consists of three steps. First, for each training set and each network structure, we train the network 100 times with different random initializations for a single epoch and record the classification results. Second, we calculate the correct classification rate (CCR) for all the samples among these 100 runs. We found that, for our data set, there are many samples consistently classified correctly or incorrectly. This result is line with the findings from [16] that, for real life data, some samples are significantly harder or easier to classify than others. In the last step, we rank the samples based on their CCR and collect the "easy" samples by taking the top $\theta\%$ samples with the highest CCR. These samples will be used as the basis for the data augmentation. Figure 2 visualizes the result of the data distillation for $\theta = 20$ via T-SNE. While the two classes strongly overlap in the original data (Fig. 2A), this overlap is greatly reduced after distillation (Fig. 2B). To quantify this effect we calculated the fraction of a data point's k nearest neighbors (according to the Euclidean distance metric, $k = 10$) that have the same class label. This number was significantly higher after distillation (median fraction 90 % in the distilled set vs. 69 % in the whole set) and this difference was statistically significant (Wilcoxon rank sum test, $p < 10^{-94}$).

Validation on MNIST. To further validate the effectiveness of the proposed distillation scheme, we also performed experiments on the well-known MNIST data set. We randomly introduced 20% uniform labeling noise (on a set of 60 000 samples), i.e., we randomly selected 20% of the samples and randomly reassigned an incorrect label. Then, we performed our proposed distillation procedure. We ran 100 single-epoch training runs and calculated the CCR of all samples. The results are shown in Fig. 2C. The black curve shows the CCRs as a function of the sample ID, where samples have been sorted by CCR. The green solid line is simply the cumulative count of samples. The green dashed line is the cumulative count of *incorrectly* labeled samples which saturated towards the right side. This result confirms that the samples

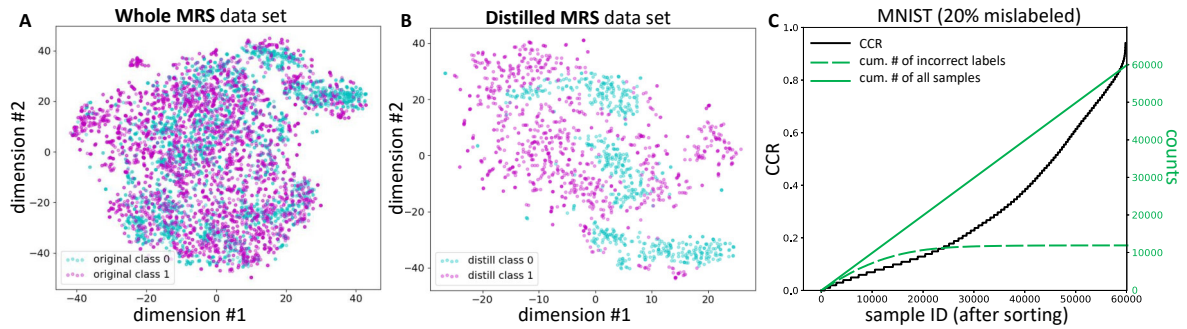


Fig. 2. Effect of data distillation. A. 2-D t-SNE visualization of all training samples. B. Same for the distilled samples (θ= 20%). C. Distillation of MNIST data with 20% noisy labels. The CCR curve is shown in black (left y-axis). The green solid line is the cumulative count of all samples and the green dashed line is the cumulative count of *incorrectly* labeled samples (right y-axis). CCR: correct classification rate

with low/high CCR are likely to be incorrectly/correctly labeled.

Data augmentation. We take the top θ% of the samples with the highest CCR, which are at the right side of the CCR curve and denote this set $\mathcal{C}$. This set forms the basis of the data augmentation. The whole data set before data augmentation is denoted $\mathcal{D}$ and the samples generated during augmentation form the set $\mathcal{A}$. Specifically, augmented sample i is created as $x_i^{\mathcal{A}} = \alpha x_j^{\mathcal{C}} + (1-\alpha)x_k^{\mathcal{D}}$, where $\alpha \in [0,1]$ is the *mixing weight*, $x_k^{\mathcal{D}} \in \mathcal{D}$ is a randomly chosen sample from the original data set that will be augmented, and $x_j^{\mathcal{C}} \in \mathcal{C}$ is a randomly chosen sample from the distilled set that will be used to augment the sample $x_k^{\mathcal{D}}$. The label of $x_i^{\mathcal{A}}$ is the same as that of $x_k^{\mathcal{D}}$.[1] The number of samples in $\mathcal{A}$ divided by that of the original set $\mathcal{D}$ is termed the augmentation factor $\Phi = |\mathcal{A}|/|\mathcal{D}|$. The full training data set is the union of the original data set and the augmentation set: $\mathcal{T} = \mathcal{D} \cup \mathcal{A}$. To deal with class imbalance, we apply a standard method of oversampling the minority class [17].

Deep Neural Network Structure. In our implementation, we adopt the residual neural network (ResNet) used in [1] and optimized parameters for our task. Specifically, we reduced the number of residual blocks to 7, which we denote as ResNet7. We increase the kernel size to 32 for 1-d convolution. The number of filters started with 16 and is increased every other block by a factor of 2. Every other block subsamples its input by a factor of 2. We apply a dropout rate of 0.55 in all blocks. We compare the full system with "ablated" versions where 1) we omit data distillation and augmentation, i.e., $\mathcal{T} = \mathcal{D}$, $\Phi = 1$, denoted as ResNet7 and 2) perform data augmentation without prior distillation, denoted as ResNet7 + DA. In this case, $\mathcal{C} = \mathcal{D}$ and augmented samples are generated by simply mixing any of the original samples as in [14].

Human vs. Machine comparison. To test how our proposed method compares to routine clinical diagnostic, a classification task on the same test set is conducted for both the network and human neuroradiologists. Eight experts with different levels of experience in 1H-MR spectroscopy (from resident to specialist of neuroradiology) were given 844 randomly selected spectra (around 105 per person). They were asked to classify each spectrum as originating from the tumor or from non-tumor tissue reviewing only the spectral lines. They were blinded to any additional information such as T2-weighted images or similar. Here, the overall performance of neuroradiologists is regarded as a collective effort. Inter-rater reliability is not applicable here, since every radiologist received a different subset of the data for classification.

4. RESULTS

To evaluate performance, we consider classification accuracy and the receiver operating characteristic (ROC) curve. We compute specificity ($\frac{\text{TN}}{\text{FP+TN}}$), sensitivity ($\frac{\text{TP}}{\text{TP+FN}}$), and the area under the ROC curve (AUC).

Training procedure. The ability of the classifier to generalize to new previously unseen patients is of great clinical importance. Therefore, we apply a 10-fold leave-subjects-out cross validation scheme. To be specific, we divide the patient list into 10 sub-lists each with around 40 patients. In each cross validation set, we withhold the data from the patients of one sub-list, while we train and validate on the data from the other sub-lists. The patient-wise accuracy is computed by averaging the classification probabilities of all voxels of that patient to obtain the final predicted label. The patient-wise accuracy is defined by the number of correct patient-wise diagnoses divided by the total number of patients in that set. The network is trained with randomly initialized weights using the Adam optimizer with default parameters and a mini-batch size of 32. The model is trained on a Linux machine with 2

[1] We also briefly experimented with a different data augmentation strategy where we reverse the role of distillation set $\mathcal{C}$ and original data set $\mathcal{D}$, i.e., we use samples from set $\mathcal{D}$ to augment samples from the distillation set $\mathcal{C}$ and keep the labels from $\mathcal{C}$ for the newly created samples. This lead to comparable results and we did not pursue the approach further.

Table 1. Network performance with the default configuration. Neuroradiologist performance was assesed on one randomly selected cross validation set, denoted as "partial" in **Data set**, whereas "whole" refers to averaging across all ten cross validation sets. Results are given as mean ± standard deviation. Accuracy is calculated in a patient-wise manner by averaging the estimated class probabilities of all voxels of a patient and thresholding (neural network) or taking the majority vote (radiologists). Dist.: distillation. DA: data augmentation.

	Data set	**Sensitivity**	**Specificity**	**AUC**	**Accuracy**
ResNet7	whole	0.65 ± 0.05	0.66 ± 0.06	0.71 ± 0.06	0.67 ± 0.05
ResNet7 + DA	whole	0.66 ± 0.06	0.67 ± 0.06	0.71 ± 0.07	0.69 ± 0.05
ResNet7 + Dist. + DA	whole	$\mathbf{0.67 \pm 0.06}$	$\mathbf{0.67 \pm 0.06}$	$\mathbf{0.72 \pm 0.09}$	$\mathbf{0.73 \pm 0.08}$
Neuroradiologists	partial	0.54	**0.88**	–	0.69
ResNet7	partial	0.62 ± 0.006	0.63 ± 0.004	0.68 ± 0.002	0.64 ± 0.005
ResNet7 + DA	partial	0.65 ± 0.003	0.65 ± 0.004	0.73 ± 0.004	0.69 ± 0.002
ResNet7 + Dist. + DA	partial	$\mathbf{0.69 \pm 0.003}$	0.69 ± 0.003	$\mathbf{0.78 \pm 0.002}$	$\mathbf{0.72 \pm 0.003}$

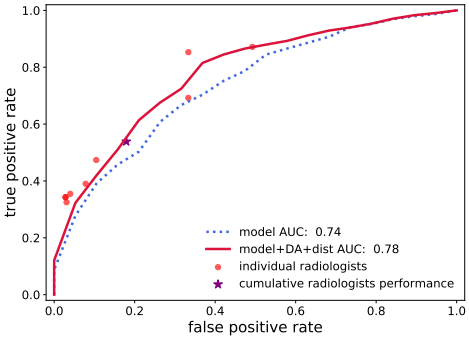


Fig. 3. Comparison of the model's and neuroradiologists' performance on one randomly selected cross validation set. The individual and collective performance of neuroradiologists are shown as red dots and purple star, respectively.

Intel(R) Xeon(R) Gold 5120 CPUs and a GeForce RTX2080ti GPU. The training requires around 10 GB of RAM and takes about 20 minutes for 200 training epochs.

We experimented with different parameters for data distillation and augmentation. Below we report the performance for the best configuration, using the parameter values highlighted in bold: $\theta = \{25, \mathbf{50}, 75\}\%$, samples from the {same class, **both classes**, opposite class} in sets $\mathcal{C}$ and $\mathcal{D}$ are mixed together with α = {0.05, 0.2, 0.35, **0.5**}, the augmentation factor is $\Phi = \{1, \mathbf{3}, 5, 9\}$.

The average results across all 10 cross-validation sets are given in Tab. 1 (upper part). It shows that our proposed method (ResNet7 + Dist. + DA) outperforms the "ablated" versions that do not utilize data augmentation (ResNet7) or use the same level of data augmentation but without prior distillation (ResNet7 + DA), similar to [14].

The lower part of Tab. 1 and Fig. 3 compare the performance of the different versions of the system to the human neuroradiologists. The standalone ResNet7 without data augmentation has an AUC of 0.68 (dotted blue line). The full framework ResNet7 + Dist. + DA achieves an AUC of 0.78. Overall, our proposed method performs on par with the group of neuroradiologists as a whole (patient-wise accuracy: 0.72 vs. 0.69). The group of radiologists exhibits greater specificity (0.88 vs. 0.69), but at the cost of lower sensitivity (0.54 vs. 0.69).

5. CONCLUSION

In this paper, we have presented a DNN-based framework which achieves performance on par with human experts on a realistic clinical task of classifying tumor and non-tumor tissues based on MRS data. We have constructed an effective data distillation and augmentation framework consisting of two steps: 1) a first neural network distills the data to alleviate label noise, 2) a data augmentation process enlarges the data set for training a second neural network for the final classification. Due to its generality, this method could be used in various other research domains. A limitation of our method is that it only takes individual spectra of a patient as input. In the future, we plan to consider patient-wise training using multiple spectra from a single individual. We hope that our framework is a step towards improving clinical practice, ultimately leading to more effective and accurate diagnosis of brain tumors in patients.

6. ACKNOWLEDGMENT

This work is supported by the China Scholarship Council (No. [2016]3100), the LOEWE Center for Personalized Translational Epilepsy Research (CePTER), and the Johanna Quandt Foundation. We thank Charles Wilmot for inspiring discussions and Marija Radović for her ideas on automating data export.

7. REFERENCES

[1] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, pp. 65, 2019.

[2] Hui Lin, Wei Zou, Taoran Li, Steven J Feigenberg, Boon-Keng K Teo, and Lei Dong, "A super-learner model for tumor motion prediction and management in radiation therapy: Development and feasibility evaluation," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[3] David Capper, David TW Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, Dominik Sturm, Christian Koelsche, Felix Sahm, Lukas Chavez, David E Reuss, et al., "Dna methylation-based classification of central nervous system tumours," *Nature*, vol. 555, no. 7697, pp. 469, 2018.

[4] Eun Kyung Park, Kwang-sig Lee, Bo Kyoung Seo, Kyu Ran Cho, Ok Hee Woo, Gil Soo Son, Hye Yoon Lee, and Young Woo Chang, "Machine learning approaches to radiogenomics of breast cancer using low-dose perfusion computed tomography: Predicting prognostic biomarkers and molecular subtypes," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[5] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[6] G Ranjith, R Parvathy, V Vikas, Kesavadas Chandrasekharan, and Suresh Nair, "Machine learning methods for the classification of gliomas: Initial results using features extracted from mr spectroscopy," *The neuroradiology journal*, vol. 28, no. 2, pp. 106–111, 2015.

[7] Nima Hatami, Michaël Sdika, and Hélène Ratiney, "Magnetic resonance spectroscopy quantification using deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 467–475.

[8] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li Jia Li, "Learning from Noisy Labels with Distillation," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 1928–1936, 2017.

[9] Kuang Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang, "CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[10] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8527–8537.

[11] Luka Smyth, Dmitry Kangin, and Nicolas Pugeault, "Training-valuenet: Data driven label noise cleaning on weakly-supervised web images," in *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2019, pp. 307–312.

[12] Riccardo Volpi, John Duchi, Hongseok Namkoong, Vittorio Murino, Ozan Sener, and Silvio Savarese, "Generalizing to unseen domains via adversarial data augmentation," *Advances in Neural Information Processing Systems*, pp. 5334–5344, 2018.

[13] Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid, "A bayesian data augmentation approach for learning deep models," in *Advances in neural information processing systems*, 2017, pp. 2797–2806.

[14] Hiroshi Inoue, "Data augmentation by pairing samples for images classification," *arXiv preprint arXiv:1801.02929*, 2018.

[15] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[16] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al., "A closer look at memorization in deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 233–242.

[17] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

Human-Expert-Level Brain Tumor Detection Using Deep Learning with Data Distillation and Augmentation

Diyuan Lu, Nenad Polomac, Iskra Gacheva, Elke Hattingen, and Jochen Triesch, *Member, IEEE*

***Abstract*—The application of Deep Learning (DL) for medical diagnosis is often hampered by two problems. First, the amount of training data may be scarce, as it is limited by the number of patients who have acquired the condition to be diagnosed. Second, the training data may be corrupted by various types of noise. Here, we study the problem of brain tumor detection from magnetic resonance spectroscopy (MRS) data, where both types of problems are prominent. To overcome these challenges, we propose a new method for training a deep neural network that distills particularly representative training examples and augments the training data by mixing these samples from one class with those from the same and other classes to create additional training samples. We demonstrate that this technique substantially improves performance, allowing our method to reach human-expert-level accuracy with just a few thousand training examples. Interestingly, the network learns to rely on features of the data that are usually ignored by human experts, suggesting new directions for future research.**

***Index Terms*—brain tumor, magnetic resonance spectroscopy, noisy labels, deep neural network, data augmentation**

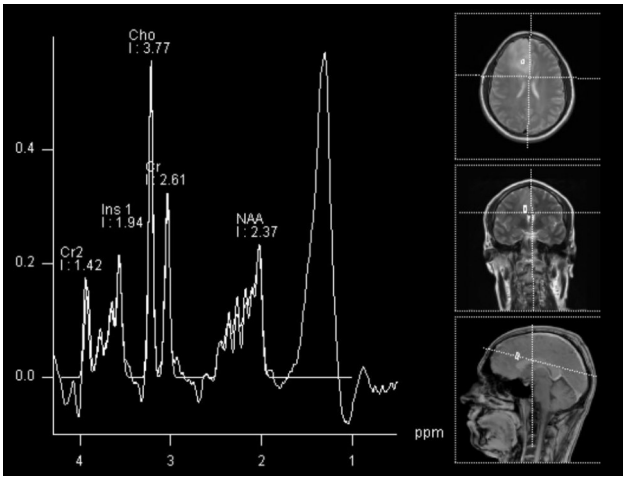


Fig. 1: Example MRS spectrum from a tumor patient

I. Introduction

MODERN machine learning (ML) approaches based on deep neural networks have recently obtained impressive results in a range of classification tasks, sometimes even outperforming human experts. These successes are made possible by the combination of 1) better learning algorithms, 2) fast, massively parallel computing hardware including graphics processing units, and 3) the availability of large training data sets. However, in many application domains, such large data sets may simply not exist or be extremely expensive to gather. This problem is particularly severe in certain medical applications, where the numbers of patients may be quite small. Typical data sets may contain only hundreds or thousands of samples, while modern ML approaches often require the estimation of many millions of free parameters. Fitting a model with many free parameters to a small set of training samples will likely lead to over-fitting and poor generalization of the learned model. This problem is aggravated if the training data are corrupted by different kinds of noise, which is often unavoidable in biomedical data.

Here, we study the problem of brain tumor detection from magnetic resonance spectroscopy (MRS) data. In clinical practice, MRS is a common tool to substantiate the diagnosis of a brain tumor and distinguish it from other medical conditions. It measures the resonant frequency shift of a chemically bound hydrogen atom (i.e., a proton), which characterizes different physiological or pathological brain metabolites. There has been increasing interest in MRS for clinical use because of the semiautomatic data acquisition, processing and quantification [1]–[3]. An example MRS spectrum from a tumor patient is shown in Fig. 1. While the interpretation of spectra is traditionally based on the size and location of certain peaks, we here use a novel approach by analysing the pattern of the MR spectrum as a whole in an unbiased fashion with machine learning.

A common problem with in-vivo MRS data is that they are quite noisy. Noise sources range from heterogeneous magnetic susceptibilities of human tissues over baseline distortions of the spectrum [1] to head movement during the procedure. Hence, the quality of spectra may be inadequate to determine precise metabolite concentrations and artefacts may resemble diagnostic features. As an additional problem, during the tissue selection process, due to the indefinable borders of gliomas, spectra from the tumor-affected hemisphere can be falsely labeled as tumor even though they contain healthy brain tissue. Furthermore, depending on the size of the selected region of interest, the number of samples collected from each patient varies substantially. Such a heterogeneous distribution of the individual training samples impedes generalization of

D. Lu and J. Triesch are with Frankfurt Institute for Advanced Studies, 60438 Frankfurt am Main, Germany. e-mail: elu, triesch@fias.uni-frankfurt.de.

N. Polomac, I. Gacheva and E. Hattingen are with the institute for Neuroradiology at Frankfurt university hospital, 60528 Frankfurt am Main, Germany. e-mail:Nenad.Polomac@kgu.de, elke.hattingen@kgu.de

the learned model — especially in a leave-one-out (LOO) cross validation scheme.

Scarcity of training data can be a big hurdle when applying DL methods to medical problems. Data augmentation is a common approach to alleviate this problem. It works by synthesizing new training data from the existing data via a variety of methods reviewed below. Here, we propose a new framework using two separate neural networks: a data distillation network to select representative training examples and a final classification network. In a nutshell, our method works by identifying data points that are "easy" to classify through the distillation network. Then, these data samples are used to synthesize a large number of new training data samples for training the final classifier. The new training samples are generated by mixing the easy samples with randomly selected data points from the same or other classes. The proposed framework is illustrated in Fig. 2. Notably, it does not require human supervision to carefully label a small data set as prototype samples [4], [5], but learns directly on the noisy labeled data. We show the benefits of this approach by demonstrating that it outperforms state-of-the-art methods and achieves human-expert-level performance. In sum, we make the following contributions:

- We propose a framework for tumor classification based on MRS data that combines deep neural networks with a novel data distillation and augmentation procedure to combat scarcity of the training data and labeling noise.
- We quantify the performance of human expert neuro-radiologists on *tumor/healthy* classification from MRS data and demonstrate that our approach achieves human-expert-level performance.
- We show that the network uses prominent features in the data that are commonly used in clinical practice, but also considers features that have not yet received much attention by medical professionals, pointing out new directions for future research.

The remainder of the paper is organized as follows. In Sect. II, we will briefly review state-of-the-art methods for data augmentation and dealing with noisy labels. In Sect. III, we present our data set and the data acquisition and preprocessing. In Sect. IV, we describe the deep neural network architecture and our new data augmentation technique. Sect. V presents and discusses our results, showing that our network can achieve human expert-level performance on this task by using the proposed data augmentation approach. Section VI concludes the paper.

II. Related work

In this section, we provide a brief review of recent research using deep neural networks in medical applications. We focus on the field of oncology and the problems of noisy labels and scarce data.

A. Deep Neural Networks

In recent years, DNN-based methods have gained more and more popularity in the healthcare domain and achieved some impressive results [2], [3], [6]–[13]. Among different network structures, Convolutional Neural Networks (CNNs) have gained great popularity [14]–[16]. They are inspired by the information processing mechanism of the visual systems of mammals where individual neurons respond to inputs in a restricted region of the visual field known as their receptive field. In comparison to fully connected neural networks, CNNs have a weight-sharing feature where neurons in different locations have identical receptive fields such that their responses can be calculated via a convolution operation. This design significantly reduces the number of trainable parameters and improves the generalization ability of the network. For example, Ng *et al.* [10] applied a Deep CNN to electrocardiography (ECG) data in heart disease classification and achieved better performance than human cardiologists. In [9], a deep CNN was trained on dermascopic melanoma detection and achieved above-dermatologist performance. [3] applied DL in Alzheimer's disease classification with MRS data. In the field of oncology, machine learning methods have obtained promising results on problems such as tumor detection, tumor segmentation, tumor progression, etc. [2], [3], [6], [7], [11]–[13]. For example, Pereira *et al.* [13] applied a deep CNN for tumor segmentation from MRI data. Podnar *et al.* [11] used a machine learning predictive model for the diagnosis of brain tumors from routine blood test results. Machine learning methods applied to MRS data, such as in [2], obtained good results in tumor grade classification according to the World Health Organization (WHO) tumor grade standard. However, learning from a larger cohort with multiple medical conditions only from MRS data has not yet been performed.

B. Learning from Noisy Labels

Noisy labels are ubiquitous in the real world. In this study, noisy labeling refers to observed labels that are incorrect, i.e., due to the labeling procedure the label assigned to the instance does not represent the class membership. Noisy labels are posing a non-trivial problem in deep model learning when an increasing ability to fit noise is accompanied with deeper layers. Given the ubiquity and importance of coping with noisy labeling, many works have been devoted to combating this problem [4], [5], [17]–[19]. One promising direction is to utilize a small set of clean labeled data [4], [17], [19], but this may not be easy to obtain. Therefore, another direction is to design models that can learn directly with noisy labels [17], [18], [20].

In [4], an auxiliary model is trained with a small but clean data set, which was manually labeled by human experts. Then the knowledge obtained by the auxiliary model is guiding the learning of the primary model in the form of one part of the primary training loss being the imitation loss of the primary model to the auxiliary trained model. Lee *et al.* [5] proposed a hybrid system, which requires a small set of representative seed instances with precise labels. Then, the automated noisy label detection is achieved with a deep CNN. Veit *et al.* [19] proposed a semi-supervised learning framework for multilabel image classification that leverages small sets of clean labels in conjunction with large amounts of noisy labels. Small sets of clean labels facilitate the learning of the mapping between

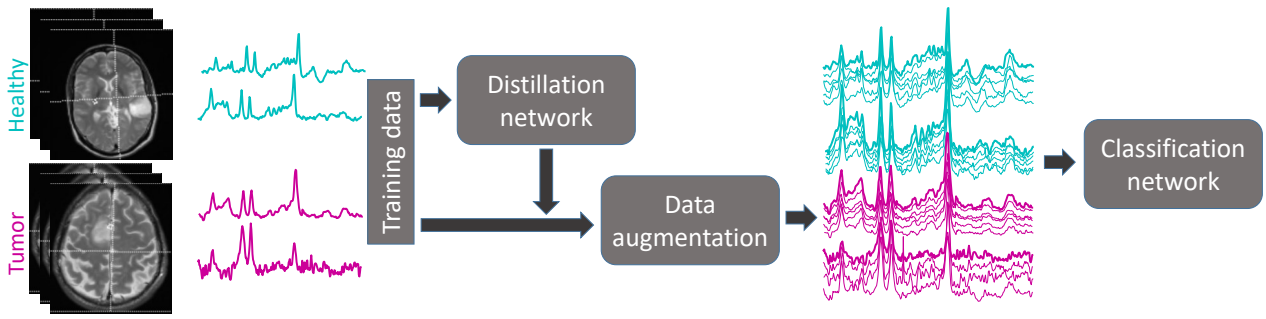


Fig. 2: Overview of the proposed approach. MRS spectra from both classes are obtained. A data distillation network automatically selects representative samples that serve as the basis for creating an augmented data set. The augmented data set is used to train a final network for classification.

noisy and clean labels, which not only reflects the noisy patterns but also the labeling structure. Han *et al.* proposed a co-teaching framework where two DNNs were trained simultaneously [17] with the whole data set. The networks train each other using small-loss training instances, since they are likely to correspond to clean annotations. The intuition is that since two neural networks are different and are equipped with different learning capabilities, during learning they may capture diverse features of the training samples and through the small-loss instances they are filtering out "clean" samples for each other. Smyth *et al.* proposed a DNN-based framework, Training-ValueNet, which evaluates the contribution of one sample to the whole learning process and then discards those that negatively contribute to the learning [18].

C. Data augmentation

A number of techniques have been explored to alleviate the problem of small training data sets. Data augmentation is a very effective way to expand the existing training set with more and diverse data in order to improve the generalization ability and incorporate invariance. Usually, data augmentation methods are domain- and dataset-specific. The fundamental rule of data augmentation is that the meaning of the target samples should be maintained regardless of the augmentation methods applied. The trained model should be reliable enough to predict the same class even when the samples are perturbed. One common class of data augmentation methods especially applicable to image data is based on different data transformations such as cropping, rotating, flipping, shearing, etc. [14], [21]. Another class of methods is referred to as adversarial training where models are trained with generated adversarial samples [22], [23]. In [22], the authors were concerned with the problem of generalizing learning from only one single source distribution to the unseen data domain. They augment the training set with generated adversarial samples. Tran *et al.* [23] proposed a joint learning scheme where a Bayesian data generator is trained with existing training samples and continuously generates new training samples for further classification. In [24], images in different styles are generated through a CycleGAN model and then used for further image classification.

In another line of thinking, data augmentation is performed by blending two or more training samples to generate new ones [25], [26]. Inoue *et al.* propose a data augmentation method by mixing randomly selected images from the training set [27]. Jaderberg *et al.* [28] presented a framework for recognizing natural scene text. In this work, a larger text corpus is generated with font rendering, creating and coloring with a background image-layer, a foreground image-layer, and an optional shadow image-layer. A natural data blending process is applied, where a random crop of an image from the training dataset is blended with each layer of the synthesized image. The three image layers are also blended together randomly to give a single output image. Summers *et al.* [26] investigated various example-mixing methods in generating new samples and found that all mixing-based data augmentation methods resulted in an improvement of baseline performance. In their work, the algorithm learned that mixing several samples of certain classes in a nonlinear way results in an improvement of the generalization ability of the learned model. However, data blending requires more delicate considerations compared to traditional data augmentation methods with various image transformations. Questions such as blending what together, how much of each component should be used, etc., need to be carefully addressed.

III. Dataset

1H-MR-spectroscopy data from 435 patients recorded in the Institute for Neuroradiology of the University Hospital, Frankfurt during the time interval from 01/2009 to 3/2019 were reviewed retrospectively. The spectroscopy was performed on a clinical 3T MR Scanner (Skyra, Siemens Medical Solutions, Erlangen, Germany) using a phased array head coil with 20 arrays and CSI-sequences with either TE = 30 ms; TR = 1500 ms; flip angle 90°; scan time of 6:11 min or TE = 135 ms; TR = 1510 ms; flip angle 90°; scan time of 3:18 min. These patients were suffering from either glial or glioneuronal first diagnosed tumors (the *tumor* group) or other non-neoplastic lesions e.g. demyelination, gliosis, focal cortical dysplasia, enlarged Virchow-Robin spaces or similar (the non-tumor/*healthy* group). The tumor group included all spectra from the tumor-affected hemisphere. The non-tumor group consisted of all spectra from both hemispheres of the patients.

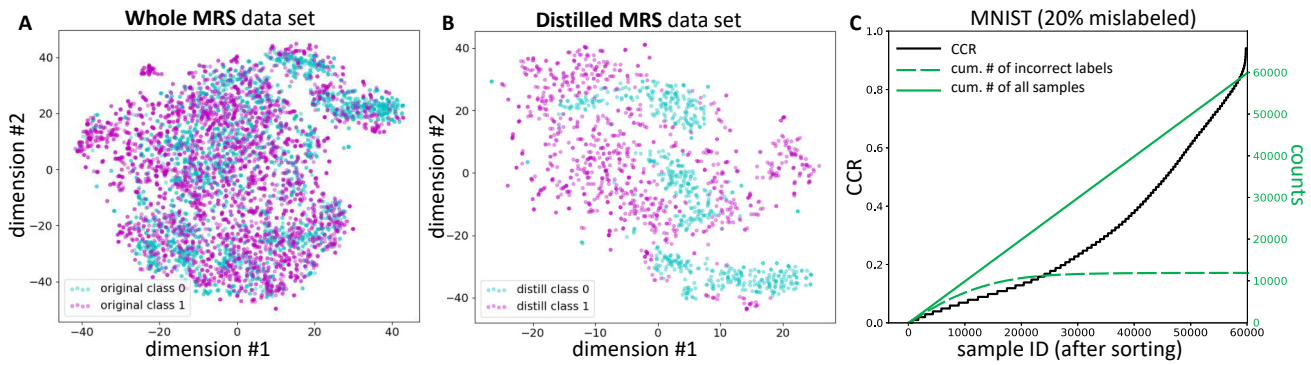


Fig. 3: Effect of data distillation. A: 2-D t-SNE visualization of all training samples. B: Same for the distilled samples ($\theta = 20\%$). C: Distillation of MNIST data with 20% noisy labels. The CCR curve is shown in black (left y-axis). The green solid line is the cumulative count of all samples and the green dashed line is the cumulative count of incorrectly labeled samples (right y-axis). CCR: correct classification rate.

As a result, 7442 spectra (3388 non-tumor and 4054 tumor) were selected for further analysis. The obtained MRS samples are saved as column vectors (288×1), shown in Fig. 1, where the y-axis shows signal intensities of different metabolites, and the x-axis represents the chemical shift positions in ppm indicating the various metabolites.

IV. Methods

In this section, we formulate our problem of classifying MRS data collected from patients with and without brain tumors into *tumor* and *healthy* classes with deep neural networks. We first outline the challenges we face in this work and propose solutions. Then, we describe in detail the network structure and the corresponding parameters. To evaluate how well our proposed method works, we construct a performance comparison in a realistic clinical setting with eight neuroradiologists.

A. Challenges

In our particular problem, we face several challenges regarding the dataset, i.e., noisy labeling, data shortage and imbalanced classes.

Noisy labeling. Infiltrative growth is an important feature of gliomas, which distinguishes them from expansively growing tumors, such as metastases. The real borders of gliomas are indefinable, which can strongly confound the selection and labeling of the voxels from multivoxel spectroscopy. One source of labeling noise is introduced when spectra from the tumor-affected hemisphere are falsely labeled as tumor-containing voxel although they contain healthy brain tissue.

Data Shortage and Class Imbalance. A large amount of training data is one of the most essential factors in training DL models successfully. However, as mentioned before, the amount of such MRS data is limited by the number of patients with the medical conditions of interest. Furthermore, as the size of the selected region of interest varies substantially for each patient, the number of samples collected from each patient also varies. Such imbalance can negatively affect the training of a classifier.

B. Proposed Solutions

Data distillation. To automatically distill the "easy" samples from the data set with noisy labeling, we propose a data distillation setup, which consists of three steps. First, for each training set and each network structure, we train 100 networks with different random initializations for a single epoch and record the classification results. Second, we calculate the correct classification rate (CCR) for all the samples among these 100 runs. We found that, for our data set, there are many samples consistently classified correctly or incorrectly. This result is in line with the findings from [29] showing that, for real life data, some samples are significantly harder or easier than others. In the last step, we rank the samples based on their CCR and select the "easy" samples by taking the top θ fraction of the samples with the highest CCR. These samples will be used as the basis for the data augmentation. Figure 3 visualizes the result of the data distillation for $\theta = 20\%$ via T-SNE. While the two classes strongly overlap in the original data (Fig. 3A), this overlap is greatly reduced after distillation (Fig. 3B). To quantify this effect we calculated the fraction of a data point's k nearest neighbors (according to the Euclidean distance metric, $k = 10$) that have the same class label. This number was significantly higher after distillation (median fraction 90 % in the distilled set vs. 69 % in the whole set) and this difference was statistically significant (Wilcoxon rank sum test, $p < 10^{-94}$).

Validation on MNIST. To further validate the effectiveness of the proposed distillation scheme, we also performed experiments on the well-known MNIST data set consisting of 60 000 hand-written digits. We randomly introduced 20% uniform labeling noise, i.e., we randomly selected 20% of the samples and randomly reassigned an incorrect label. Then, we performed our proposed distillation procedure. We ran 100 single-epoch training runs and calculated the CCR of all samples. The results are shown in Fig. 3C. The black curve shows the CCR as a function of the sample ID, where samples have been sorted by CCR. The green solid line is the cumulative count of total samples, which is a straight line that starts from (0, 0) and ends at (60 000, 60 000).

The green dashed line is the cumulative count of incorrectly labeled samples. We can see that the samples with incorrect labels populate the left part of the CCR curve. This result confirms that the samples with low/high CCR are likely to be incorrectly/correctly labeled.

To investigate the effect of our data distillation scheme on training, we performed the following experiment. We first trained a network on MNIST with clean labels. The test accuracy of this baseline was 96.1%. Then we trained a network with 20% noisy labels. The validation accuracy of this network was 73.2% and when we tested it on a test set with clean labels, its accuracy was 92.0%, i.e., 4.1% worse than the baseline without labeling noise. Finally, we used our data distillation and selected only training samples corresponding to the top 80% highest CCR from the training set with noisy labels to train a new network. When we tested this model on the test set with clean labels it achieved a test accuracy of 95.6%, i.e., the distillation process had largely reverted the harmful effect of the noisy labels. This shows that our data distillation procedure can successfully combat the adverse effects of labeling noise.

Data augmentation. Returning to the problem of tumor classification, we take the fraction of θ samples at the right side of the CCR curve and denote this set $\mathcal{C}$. This set forms the basis of the data augmentation. The whole data set before data augmentation is denoted $\mathcal{D}$ and the samples generated during augmentation form the set $\mathcal{A}$. Specifically, augmented sample i is created as

$$x_i^{\mathcal{A}} = \alpha x_j^{\mathcal{C}} + (1-\alpha) x_k^{\mathcal{D}}, \tag{1}$$

where $\alpha \in [0,1]$ is the *mixing weight*, $x_j^{\mathcal{C}} \in \mathcal{C}$ is a randomly chosen sample from the distilled set that will be augmented, and $x_k^{\mathcal{D}} \in \mathcal{D}$ is a randomly chosen sample from the original data set that is mixed with the distilled sample. The label of $x_i^{\mathcal{A}}$ is the same as that of $x_j^{\mathcal{D}}$. The number of samples in $\mathcal{A}$ divided by that of the original set $\mathcal{D}$ is termed the augmentation factor $\Phi = |\mathcal{A}|/|\mathcal{D}|$. The full training data set is the union of the original data set and the augmentation set: $\mathcal{T} = \mathcal{D} \cup \mathcal{A}$. We propose three different augmentation strategies: augment with the same class (aug-with-same), augment with the opposite class (aug-with-other) and augment with both classes (aug-with-both). Based on the choice of the augmentation strategy, $x_k^{\mathcal{D}}$ could be randomly selected from either class groups or both. To deal with the class imbalance, we apply the method of oversampling the minority class described in [30].

C. Deep Neural Network Structure

In our implementation, we apply the residual neural network proposed by He *et al.* as the backbone [16]. Residual neural networks feature skip-connections, which connect the input of one layer and the pre-activation of another layer skipping multiple layers in between. This structure is usually termed a residual block. One block usually consists of multiple computational layers such as convolutional or dense layers with batch normalization [31], drop-out [32], and a non-linear activation transformation [33]. The input to the residual block is split into two branches: the main branch with convolution or dense matrix multiplication, batch-normalization, drop-out and the other branch usually with the identity transformation or max-pooling. The combination of the outputs of these two branches is passed through a non-linear activation function as the input of the next block. We implement a deep residual neural network with 8 residual blocks following the classic structure from [16], including 17 convolutional layers and skip connections. It is inspired by the network architecture in [10]. Each residual block consists of two convolutional layers with batch normalization, drop out and ReLU non-linear activation functions. The convolutional layers have a filter width of 32×1. Experimenting with different kernel sizes, 32 gives good performance. The number of filters increases by a factor of 2 in every other block starting from 16. There is a sub-sample layer of factor 2 in every other block occurring at the same time when increasing the number of filters. We apply a dropout rate of 0.55 in all blocks. A global average pooling (GAP) layer follows the last convolutional layer to provide further visualization, which is termed a class activation map (CAM) [34]. The GAP layer is followed by a soft-max layer, which outputs a probability distribution over the two possible classes. The detailed parameters of the network structure are shown in Table I.

TABLE I: Proposed network structure. The **Config** column shows the configuration in convolutional and dense layers (filter size 32×1 and the number of filters, or the number of units in the dense layer). The number of filters is increased every other block by a factor of 2. Every other block subsamples its input by a factor of 2, indicated by the value of **Stride**. Here, the batch size at the first dimension is omitted in the output shape column. GAP: global average pooling.

Name	Config	Stride	Output size
Conv	$\begin{bmatrix} 32\times 1, & 16 \end{bmatrix}$	1	[batch size, 288, 1, 16]
ResBlock 1	$\begin{bmatrix} 32\times 1, & 16 \\ 32\times 1, & 16 \end{bmatrix}$	1	[batch size, 144, 1, 16]
ResBlock 2	$\begin{bmatrix} 32\times 1, & 16 \\ 32\times 1, & 16 \end{bmatrix}$	1	[batch size, 144, 1, 16]
ResBlock 3	$\begin{bmatrix} 32\times 1, & 32 \\ 32\times 1, & 32 \end{bmatrix}$	2	[batch size, 72, 1, 32]
ResBlock 4	$\begin{bmatrix} 32\times 1, & 32 \\ 32\times 1, & 32 \end{bmatrix}$	1	[batch size, 72, 1, 32]
ResBlock 5	$\begin{bmatrix} 32\times 1, & 64 \\ 32\times 1, & 64 \end{bmatrix}$	2	[batch size, 36, 1, 64]
ResBlock 6	$\begin{bmatrix} 32\times 1, & 64 \\ 32\times 1, & 64 \end{bmatrix}$	1	[batch size, 36, 1, 64]
ResBlock 7	$\begin{bmatrix} 32\times 1, & 128 \\ 32\times 1, & 128 \end{bmatrix}$	2	[batch size, 18, 1, 128]
ResBlock 8	$\begin{bmatrix} 32\times 1, & 128 \\ 32\times 1, & 128 \end{bmatrix}$	1	[batch size, 18, 1, 128]
GAP			[batch size, 128]
Dense	2		[batch size, 2]

D. Visualization through Class Activation Maps

Modern DL techniques are often viewed as black-box methods, where the decision making process is difficult to understand for humans. It raises worrying questions and hinders the practical deployment of such techniques. Much effort has been devoted to develop explainable and interpretable DL approaches [34]–[37].

In our work, we apply a GAP layer to reduce the risk of over-fitting and provide further visualization of the network decision making processes. The GAP squashes the output of each feature map with the shape $h \times w \times d$ from the previous layer into one single value with the shape of $1 \times 1 \times d$ reducing the number of features by $h \times w$ fold. The output of the GAP layer is fed directly to the final classification layer. Intuitively, the GAP operation converts feature maps into weights that represent the "importance" of all feature maps, namely the CAMs. An added value of this method is that we can easily trace back the "importance" to the input space and visualize how much of each part of the input contributes to the final classification decision.

E. Quantifying Performance

To illustrate how well our proposed method works in comparison to routine clinical diagnostic, a classification task on the same test set is conducted for both the network and human neuroradiologists. Eight experts with different levels of experience in the 1H-MR spectroscopy (from resident to specialist of neuroradiology), were given 844 randomly selected spectra (around 105 per person). They were asked to classify each spectrum as originating from the tumor or from non-tumor tissue reviewing only the spectral lines. They were blinded to any additional information such as T2-weighted images or similar. The overall performance of neuroradiologists is regarded as a collective effort. Inter-rater reliability is not applicable here, since every radiologist received different subsets of the data to classify.

To evaluate performance, we use the receiver operating characteristic (ROC) curve, which is a gold standard to evaluate the discriminative ability of a classifier. It is constructed by varying the classification threshold and calculating the true positive (TP), false positive (FP), true negative (TN), and false negative (FN). We report sensitivity = $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$, specificity = $\frac{\mathrm{TN}}{\mathrm{TN+FP}}$, area under the ROC curve (AUC), accuracy, F1-score = $\frac{2\mathrm{TP}}{2\mathrm{TP+FP+FN}}$, and Matthews correlation coefficient (MCC) = $\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP+FP})(\mathrm{TP+FN})(\mathrm{TN+FP})(\mathrm{TN+FN})}}$. The area under the curve (AUC) is a scalar value between zero and one which characterizes the goodness of the classifier. The MCC is generally considered as a balanced measure which takes into TN, TN, TN, and TN, and it can be used even if the classes are not balanced. We also compare our results with three baseline methods: a fully-connected network (FNN), a recurrent neural network (RNN), and an Inception network. To investigate the effect of our proposed data augmentation and data distillation methods, we also report the performance with ablation for all network structures.

V. Results

A. Training procedure

The ability of the classifier to generalize to new previously unseen patients is of great clinical importance. Therefore, we apply a 10-fold leave-subjects-out cross validation scheme. To be specific, we divide the patient list into 10 sub-lists each with around 40 patients. In each cross validation set, we withhold the data from the patients of one sub-list, while we train and validate on the data from the other sub-lists. The patient-wise accuracy is computed in each leave-out test set. For each patient, the classification probability of all voxels are averaged to get the probabilities of each class. Then, the patient-wise diagnosis is obtained as the class that has the highest probability. The patient-wise accuracy is defined by the number of correct patient-wise diagnoses divided by the total number of patients in that set. We randomly select one cross validation set which consists of 844 spectra from 40 patients for the final test against human neuroradiologists. The network is trained with randomly initialized weights using the Adam optimizer with default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a mini-batch size of 32. The model is trained on a Linux machine with 2 Intel(R) Xeon(R) Gold 5120 CPUs and a GeForce RTX2080ti GPU. The training does not require more than 10GB RAM. It takes 20 minutes to finish 200 epochs of training.

To get an average performance of the effect of the proposed distillation process, we train the whole framework twice, i.e., a distillation network, which collects certain samples and a primary classifier with proposed data augmentation on all 10 cross-validation sets, with different random seeds. The results are averaged across the 10 cross-validation sets as well as the two runs. The overall performance of different baseline models with and without our proposed distillation and data augmentation is reported in Table. II. It shows that our proposed method slightly outperforms the human neuroradiologists. The performance comparison between different baseline models and neuroradiologists is shown in Table. III

We experimented with different parameters in distillation and data augmentation, and here we report the performance under the best configuration, highlighted in bold: $\theta = \{25, \mathbf{50}, 75\}\%$, samples from the {same class, **both classes**, opposite class} in sets $\mathcal{C}$ and $\mathcal{D}$ are mixed together with α = {0.05, 0.2, 0.35, **0.5**}, the augmentation factor is $\Phi = \{1, \mathbf{3}, 5, 9\}$.

B. Data Distillation

To further illustrate the effect of the proposed distillation on our MRS data set, we randomly selected 500 samples that the network assigns high probability while doing classification from one training set and obtained expert annotation of them. The experts were asked to validate whether the samples are correctly labeled based on the tissue's location. For example, a sample, labeled as tumor since it is from a tumor patient, resides outside the tumor tissue would account for an incorrect labeling. We expect that the correctly labeled tissue based on the MRS characteristic and the location would appear towards the right side of the CCR curve, illustrated in Fig 3C. One

TABLE II: Performance measures with default configurations ($\Phi = 3, \alpha = 0.5$ and augment with both classes). The performance is averaged across all ten cross validation sets. Results are given as mean $\pm$ standard deviation. The best performance is in **bold**. Accuracy is calculated in a patient-wise manner by averaging the estimated class probabilities of all voxels of a patient and thresholding the result. Dist.: distillation. DA: data augmentation. MCC: Matthews correlation coefficient

	Sensitivity	**Specificity**	**AUC**	**Accuracy**	**F1-score**	**MCC**
FNN	0.63 ± 0.05	0.64 ± 0.05	0.68 ± 0.07	0.66 ± 0.06	0.65 ± 0.08	0.26 ± 0.10
FNN + DA	0.66 ± 0.05	0.66 ± 0.05	0.71 ± 0.07	0.69 ± 0.04	0.65 ± 0.09	0.31 ± 0.09
FNN + Dist. + DA	0.67 ± 0.02	0.66 ± 0.02	0.72 ± 0.04	0.71 ± 0.03	0.68 ± 0.08	0.32 ± 0.04
FNN + Dist. + DA (new)	0.66 ± 0.05	0.66 ± 0.06	0.72 ± 0.06	0.69 ± 0.08	0.66 ± 0.09	0.30 ± 0.11
Inception	0.64 ± 0.04	0.62 ± 0.07	0.65 ± 0.06	0.69 ± 0.06	0.64 ± 0.08	0.25 ± 0.10
Inception + DA	0.66 ± 0.06	0.66 ± 0.06	0.70 ± 0.04	0.68 ± 0.08	0.64 ± 0.12	0.31 ± 0.13
Inception + Dist. + DA	0.66 ± 0.04	0.67 ± 0.04	0.71 ± 0.05	0.69 ± 0.07	0.65 ± 0.07	0.31 ± 0.08
Inception + Dist. + DA (new)	0.68 ± 0.05	0.65 ± 0.08	0.70 ± 0.06	0.72 ± 0.06	0.68 ± 0.08	0.32 ± 0.05
RNN	0.64 ± 0.05	0.65 ± 0.06	0.69 ± 0.06	0.66 ± 0.07	0.65 ± 0.08	0.28 ± 0.11
RNN + DA	0.66 ± 0.04	0.66 ± 0.04	0.71 ± 0.07	0.67 ± 0.07	0.67 ± 0.08	0.30 ± 0.08
RNN + Dist. + DA	0.65 ± 0.04	0.68 ± 0.08	0.72 ± 0.07	0.68 ± 0.07	0.67 ± 0.07	0.32 ± 0.11
ResNet7	0.65 ± 0.05	0.66 ± 0.06	0.71 ± 0.06	0.67 ± 0.05	0.66 ± 0.08	0.29 ± 0.05
ResNet7 + DA	0.66 ± 0.06	0.67 ± 0.06	0.71 ± 0.07	0.69 ± 0.05	0.67 ± 0.07	0.29 ± 0.11
ResNet7 + Dist. + DA	$\mathbf{0.67 \pm 0.06}$	0.67 ± 0.06	$\mathbf{0.72 \pm 0.09}$	$\mathbf{0.73 \pm 0.08}$	$\mathbf{0.68} \pm 0.07$	$\mathbf{0.32} \pm 0.13$

TABLE III: Performance comparison with Neuroradiologists. The performance of the neuroradiologists is computed on one randomly selected cross validation set. Results are given as mean $\pm$ standard deviation. Accuracy is calculated in a patient-wise manner by averaging the estimated class probabilities of all voxels of a patient and thresholding (neural network) or taking the majority vote (radiologists). Dist.: distillation. DA: data augmentation. MCC: Matthews correlation coefficient

	Sensitivity	**Specificity**	**AUC**	**Accuracy**	**F1-score**	**MCC**
Neuroradiologists	0.54	**0.88**	–	0.69	0.56	0.58
ResNet7 + Dist. + DA	$\mathbf{0.69 \pm 0.004}$	0.69 ± 0.002	$\mathbf{0.75 \pm 0.001}$	$\mathbf{0.72 \pm 0.001}$	$\mathbf{0.62 \pm 0.001}$	$\mathbf{0.37 \pm 0.003}$

simple question we can ask is that based on the CCR we obtained about those expert-validated samples, what is the optimal cut-off threshold of CCR such that we can distinguish the correctly- and wrongly-labeled samples from the original labeling process. To answer this question, we did a ROC analysis where the expert-label is 1 when the sample is correctly labeled and 0, otherwise. The target scores used to compute the ROC AUC curve are their corresponding CCR. We found that the optimal cut-off CCR threshold is 0.49, which is at 55-th percentile. This finding is consist with our empirical choice that θ = **50**% works the best among {25, 50, 75}%.

C. Data Augmentation

In this section, we discuss different effects on learning resulting from different options including the mixing weight α, the augmentation factor, the index of the last source epoch from which we collect the certain samples and the three augmentation strategies (aug-with-same, aug-with-other and aug-with-both). We measure the AUC of the ROC curve with different parameter options for different augmentation strategies. The results are averaged across all 10 cross-validation sets with two different initial distillation networks.

Adding noise to augment data is a common practice in image data enrichment. Here, we also report results for the case when Gaussian noise is added to augment the data (noise augmentation). We explore different hyper-parameters such as noise amplitude and augmentation factor, and report the performance under the parameters that yielded the best result during the exploration.

In Fig. 4, we show the performance under different configurations under different augmenting methods. We note a number of observations. First, the aug-with-both method with mixing weight α = 0.5, augmentation factor Φ = 3 yields the best performance. Second, in aug-with-same and aug-with-both cases, with an increasing α, the performance increases in almost all Φ cases. The However, the aug-with-other show exactly the opposite trend. Third, with a small α, the aug-with-other method is very robust and the performance is relatively insensitive to the change of Φ.

D. Human vs. Machine

To assess how well our proposed method works in a more realistic clinical setting, we compared it to human neuroradiologists on one randomly selected test set. The result is shown in Fig. 5. The test set is divided into eight subsets and assigned to eight neuroradiologists. The performance of each individual neuroradiologists is denoted as a red dot, the collective performance is shown as a purple diamond. The performance of the model is computed in each corresponding subset as each individual neuroradiologist and averaged across all subsets. The model without data augmentation has an AUC of 0.72 (dashed blue line), a MCC of 0.27, and an F1-score of 0.56. The model with data distillation and augmentation achieves an AUC of 0.77 (solid orange), a MCC of 0.37, and an F1-score of 0.62. It shows that the performance of our proposed method is on par with the group of neuroradiologists

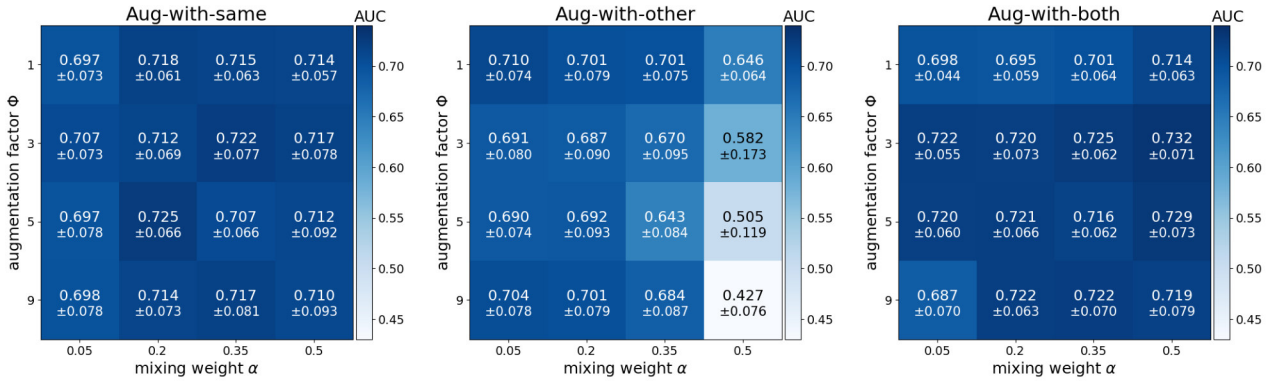


Fig. 4: Performance averaged among all test sets with various augmentation parameters (mixing weight α = {0.05, 0.2, 0.35, 0.5}, the augmentation factor $\Phi = \{1, 3, 5, 9\}$) for the different augmentation methods.

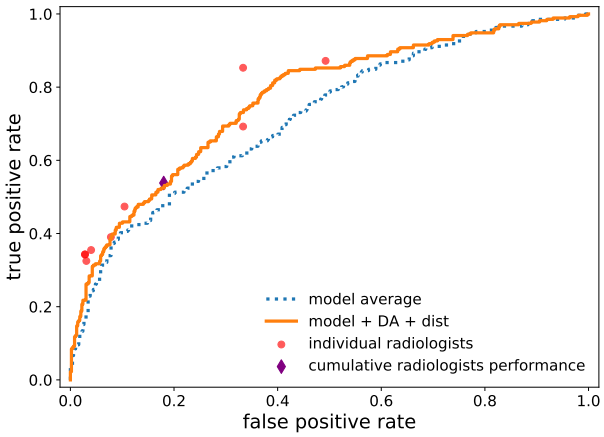


Fig. 5: Comparison of the proposed model and neuroradiologists on one randomly selected cross validation set. The individual and collective performance of neuroradiologists are shown as red dots and purple diamond, respectively. The average ROC curve of the plain ResCNN model and our proposed model with default augmentation parameters (aug-with-both method, augmentation factor Φ = 3 and the mixing weight is $\alpha = 0.5$) are depicted in dashed and solid lines, respectively.)

(sensitivity 0.69 vs. 0.54, specificity 0.69 vs. 0.88, accuracy: 0.72 vs. 0.69, F1-score: 0.62 vs. 0.56, and MCC: 0.37 vs. 0.58).

E. Feature Visualization

As described in section IV, we apply a GAP layer after the convolutional layers to prevent over-fitting and benefit from the possibility of visualizing class activation maps. These show how the network is making the final decision by assigning different weights, which can be interpreted as "importance", to different regions in the input data.

In Fig. 6, we show some examples of CAMs with original MRS samples from both classes. The results show that the CAMs vary with regard to specific samples. To interpret these CAMs, one must not only focus on the highest peak but

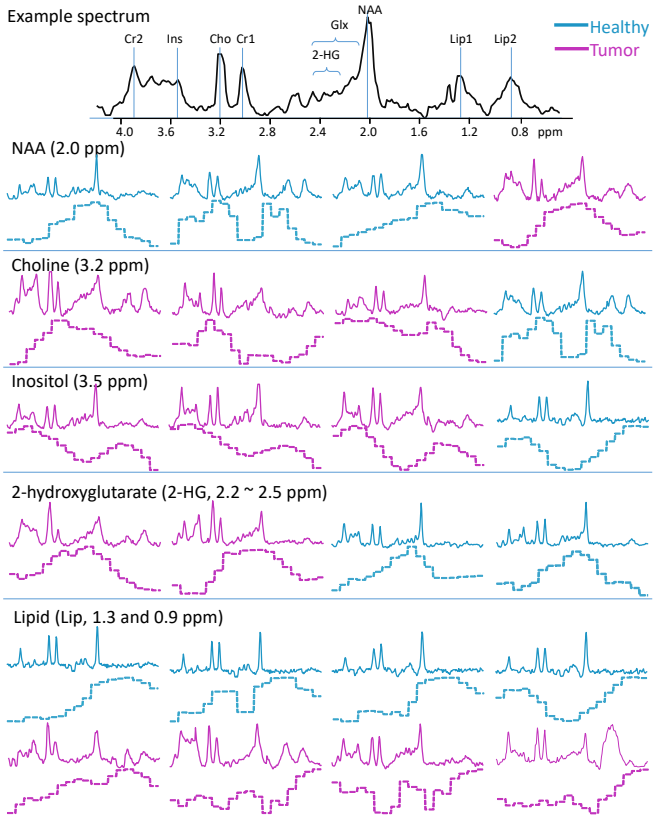


Fig. 6: Class activation maps of examples from both classes during test. At the top is an example where various metabolite peaks have been marked. Examples from class *healthy* and *tumor* are color coded in green and purple, respectively. Solid lines are original examples and the dashed line below is the corresponding CAM. Cr1, Cr2: creatine, Ins: myo-inositol. Cho: choline. NAA: n-acetylaspartic acid. Glx: glutamine, Lip: lipid

rather the overall shape together with the original spectra. Note that the "importance" does not reflect whether the signal intensity of the corresponding metabolite is high or low. The co-occurrence of high "importance" regions provides insights

in the CAM interpretation. We can see that the network considers various common metabolites during the classification. Interestingly, for the *healthy* class the network also pays more attention to the plateau left of the dominant NAA peak. In the *tumor* spectra, this part of the spectrum appears as a rising slope, and represents the oncometabolite 2-hydroxyglutatat [38], [39] as well as tumor associated metabolites like glutamine [40]. The Ins peak together with Cr2 and Lip regions are highly interesting. A high Ins peak with above-baseline Lip peaks highly suggests tumor presence and a low Ins concentration with almost no free lipids suggests the *healthy* class [41]. In cases where a high "importance" is assigned to the Cho region, the *tumor* spectra show a high Cho peak flanked by other tumor-associated metabolite peaks (glycine, myo-inositol) [42], [43]. On the other hand, the *healthy* group shows a similar or smaller Cho peak as the Cr1 peak.

In the Appendix and in Fig. 7, we discuss additional insights from unsupervised k-means clustering of the tumor spectra.

VI. Conclusion

In this paper, we present a DNN-based framework, which achieves above human-level performance on a realistic clinical task of classifying tumor and non-tumor tissues based on MRS data. We construct an effective data cleaning and augmentation framework consisting of two steps: 1) a data distillation network to clean noisy labeled data, 2) a data augmentation process, which enlarges the data set acquired in the first step for training a primary neural network for the final classification. Due to its generality, this data augmentation method could be used in various other research domains. By exploring various configurations of the proposed data augmentation method, we further demonstrate that data augmentation by mixing samples from both classes is more stable and yields better results. A deep residual neural network is used as the primary learning model and a global average pooling (GAP) layer at the end of all convolutional layers provides us with a visualization of how much each part of the input contributes to the final classification decision. Our proposed framework outperforms neuroradiologists on sensitivity and patient-wise diagnosis accuracy with an area under the ROC curves of 0.77. With an improved capability of coping with noisy labeling and the scarcity of the training data, we believe that the framework proposed in this work could improve clinical practice, ultimately leading to more effective and accurate diagnosis of brain tumors in patients.

Acknowledgment

This work is supported by the China Scholarship Council (No. [2016]3100), the LOEWE Center for Personalized Translational Epilepsy Research (CePTER), and the Johanna Quandt Foundation. Special thanks to Charles Wilmot for inspiring discussions. Furthermore a particular appreciation goes to Marija Radović for her ideas on automating the data export.

Appendix

To get an overview of the data we use in this task, we performed k-means clustering on the whole data set $\mathcal{D}$, which has 7442 spectra (3388 healthy and 4054 tumor). The euclidean distance is used as the criterion to cluster the data. The number of clusters is determined by the elbow point in the inertia curve [44], where the within cluster distance does not decrease significantly with an increasing number of clusters (a number of seven is chosen in this study). We did find a large overlap between two classes as we expected.

The clustering results are shown in Fig. 7. The cross-tab relation, which is a frequency count of one variable (*healthy* or *tumor*) in each cluster is shown in Fig. 7. A. For example, cluster 1 contains 18.5% of the healthy spectra and 8.7% of the tumor spectra. We can see that 1) there are samples from healthy and tumor group in every cluster, 2) there are roughly equal amounts of healthy and tumor samples in clusters 2, 3, 5, 6 and 7, 3) the majority of samples in cluster 1 are showing typical features of *healthy* (no Lip1 or Lip2 concentration [45]) and those of cluster 4 are mainly typical *tumor* (high Lip peaks, an elevated Cho peak, high Glx region, etc. [46]), and 4) the majority of the spectra are neither typical healthy nor tumor, rather somewhere in between. The positions of typical metabolites are demonstrated in Fig. 6-A. The mean spectra of those clusters illustrate commonly applied clinical assessment criteria: in healthy tissues, there is a dominant peak at NAA and almost no mobile lipids to be detected since they are mostly confined to the membrane [46]. In tumor tissues, there are elevated Cho and Lip peaks [45]. A median to high Cho peak with easily visible Cr peaks can contribute to the identification of a tumor [45]. The clustering results support our argument that the labeling process is noisy, so the spectra from both classes are largely mixed with each other.

References

[1] Roland Kreis. Issues of spectral quality in clinical 1h-magnetic resonance spectroscopy and a gallery of artifacts. *NMR in Biomedicine*, 17(6):361–381, 2004.

[2] G Ranjith, R Parvathy, V Vikas, Kesavadas Chandrasekharan, and Suresh Nair. Machine learning methods for the classification of gliomas: Initial results using features extracted from mr spectroscopy. *The neuroradiology journal*, 28(2):106–111, 2015.

[3] Cristian R Munteanu, Carlos Fernandez-Lozano, Virginia Mato Abad, Salvador Pita Fernández, Juan Álvarez-Linera, Juan Antonio Hernández-Tamames, and Alejandro Pazos. Classification of mild cognitive impairment and alzheimer's disease with machine-learning techniques using 1h magnetic resonance spectroscopy data. *Expert Systems with Applications*, 42(15-16):6205–6214, 2015.

[4] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li Jia Li. Learning from Noisy Labels with Distillation. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:1928–1936, 2017.

[5] Kuang Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[6] Hui Lin, Wei Zou, Taoran Li, Steven J Feigenberg, Boon-Keng K Teo, and Lei Dong. A super-learner model for tumor motion prediction and management in radiation therapy: Development and feasibility evaluation. *Scientific reports*, 9(1):1–11, 2019.

[7] Eun Kyung Park, Kwang-sig Lee, Bo Kyoung Seo, Kyu Ran Cho, Ok Hee Woo, Gil Soo Son, Hye Yoon Lee, and Young Woo Chang. Machine learning approaches to radiogenomics of breast cancer using low-dose perfusion computed tomography: Predicting prognostic biomarkers and molecular subtypes. *Scientific reports*, 9(1):1–11, 2019.

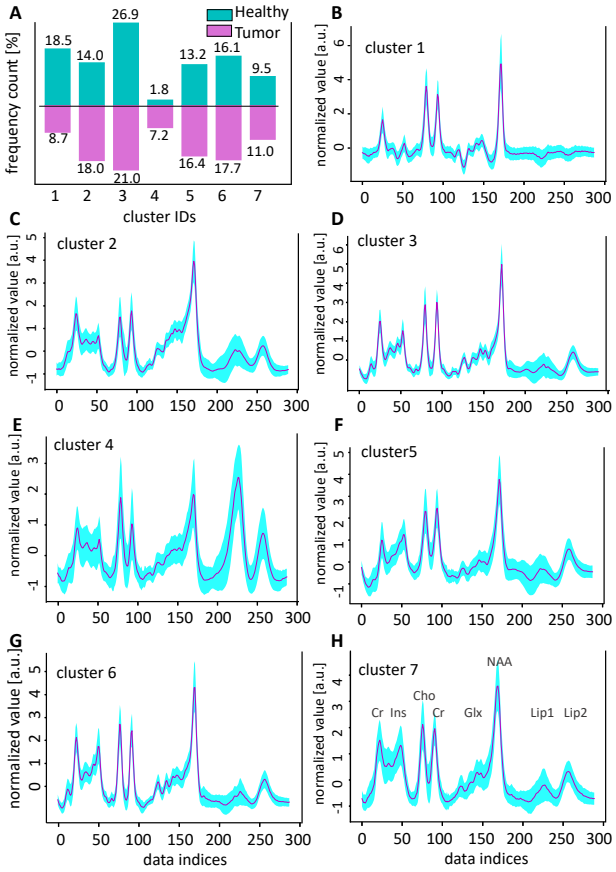


Fig. 7: k-means clustering on the whole data set. **A**. Cross-tab relation of the clustering results. **B-H**. Mean spectra of each cluster (magenta) with standard deviation (cyan).

[8] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[9] Holger A Haenssle, Christine Fink, R Schneiderbauer, Ferdinand Toberer, Timo Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, L Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018.

[10] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65, 2019.

[11] Simon Podnar, Matjaž Kukar, Gregor Gunčar, Mateja Notar, Nina Gošnjak, and Marko Notar. Diagnosing brain tumours by routine blood tests using machine learning. *Scientific reports*, 9(1):1–7, 2019.

[12] David Capper, David TW Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, Dominik Sturm, Christian Koelsche, Felix Sahm, Lukas Chavez, David E Reuss, et al. Dna methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469, 2018.

[13] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

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

[15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[16] 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.

[17] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, 2018.

[18] Luka Smyth, Dmitry Kangin, and Nicolas Pugeault. Training-valuenet: Data driven label noise cleaning on weakly-supervised web images. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 307–312. IEEE, 2019.

[19] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. 2017.

[20] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. 2017.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84–90, 2017.

[22] Riccardo Volpi, John Duchi, Hongseok Namkoong, Vittorio Murino, Ozan Sener, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in Neural Information Processing Systems*, pages 5334–5344, 2018.

[23] Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid. A bayesian data augmentation approach for learning deep models. In *Advances in neural information processing systems*, pages 2797–2806, 2017.

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

[25] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[26] Cecilia Summers and Michael J Dinneen. Improved mixed-example data augmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1262–1270. IEEE, 2019.

[27] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018.

[28] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. pages 1–10, 2014.

[29] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 233–242. JMLR. org, 2017.

[30] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

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

[32] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[33] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, page 3, 2013.

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 2921–2929, 2016.

[35] Pieter Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and Patternattribution. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pages 1–12, 2018.

[36] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding Neural Networks Through Deep Visualization. *Computer Science*, 2015.

[37] Bach Sebastian, Binder Alexander, Montavon Grégoire, Klauschen Frederick, Müller Klaus-Robert, Samek Wojciech, and Suarez Oscar Deniz. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *Plos One*, 10(7):e0130140–.

[38] O. C. Andronesi, O. Rapalino, E. Gerstner, A. Chi, T. T. Batchelor, D. P. Cahill, A. G. Sorensen, and B. R. Rosen. Detection of oncogenic IDH1 mutations using magnetic resonance spectroscopy of 2-hydroxyglutarate. *J. Clin. Invest.*, 123(9):3659–3663, Sep 2013.

[39] K. E. Yen, M. A. Bittinger, S. M. Su, and V. R. Fantin. Cancer-associated IDH mutations: biomarker and therapeutic opportunities. *Oncogene*, 29(49):6409–6417, Dec 2010.

[40] Y. Li, P. Larson, A. P. Chen, J. M. Lupo, E. Ozhinsky, D. Kelley, S. M. Chang, and S. J. Nelson. Short-echo three-dimensional H-1 MR spectroscopic imaging of patients with glioma at 7 Tesla for characterization of differences in metabolite levels. *J Magn Reson Imaging*, 41(5):1332–1341, May 2015.

[41] A. C. Kuesel, K. M. Briere, W. C. Halliday, G. R. Sutherland, S. M. Donnelly, and I. C. Smith. Mobile lipid accumulation in necrotic tissue of high grade astrocytomas. *Anticancer Res.*, 16(3B):1485–1489, 1996.

[42] W. Moller-Hartmann, S. Herminghaus, T. Krings, G. Marquardt, H. Lanfermann, U. Pilatus, and F. E. Zanella. Clinical application of proton magnetic resonance spectroscopy in the diagnosis of intracranial mass lesions. *Neuroradiology*, 44(5):371–381, May 2002.

[43] E. Hattingen, H. Lanfermann, J. Quick, K. Franz, F. E. Zanella, and U. Pilatus. 1H MR spectroscopic imaging with short and long echo time to discriminate glycine in glial tumours. *MAGMA*, 22(1):33–41, Feb 2009.

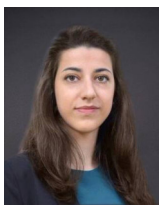
[44] Trupti M Kodinariya and Prashant R Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[45] Guoguang Fan. Magnetic resonance spectroscopy and gliomas. *Cancer Imaging*, 6(1):113–115, 2006.

[46] C. Rae. Re: Magnetic resonance spectroscopy of the brain: review of metabolites and clinical applications. *Clinical Radiology*, 64(10):0–1043.

Diyuan Lu received the B.Sc. and M.Sc. degrees from Northwestern Polytechnical University, Xi'an, Shaanxi, China, in 2012 and 2016, respectively. She is currently pursuing the Ph.D. degree at Frankfurt Institute for Advanced Studies, Goethe University, Frankfurt am Main, Germany. Her current research interests include deep learning and machine learning with medical applications.

Nenad Polomac gradated as the MD at the Medical Faculty of Belgrade, Serbia in 2009. Since 2017 resident at the institute for Neuroradiology at Frankfurt university hospital. His research interest is quantitative neuroimaging and application of machine learning algorithms on this data.



Iskra Gacheva graduated in medicine at the University of Frankfurt in 2019. Since then she is a doctoral candidate at the institute for Neuroradiology at Frankfurt university hospital. Her research interest is the application of machine learning algorithms in metabolic neuroimaging.

Elke Hattingen graduated as Dr. med. by the Faculty of Medicine of the University of Freiburg in 1994. In 2009, she received the postdoctoral lecturing qualification (habilitation). From 2014 to 2017, she was a professor of Neuroradiology in Bonn University. From 2018 until now, she is a professor of Neuroradiology at Frankfurt university. Her current research interest is in metabolic and quantitative imaging of brain tumors, epilepsy, multiple sclerosis and cerebrovascular diseases. Future projects will focus on radionomics.

Jochen Triesch received his Diploma and Ph.D. degrees in Physics from the University of Bochum, Germany, in 1994 and 1999, respectively. After two years as a post-doctoral fellow at the Computer Science Department of the University of Rochester, NY, USA, he joined the faculty of the Cognitive Science Department at UC San Diego, USA as an Assistant Professor in 2001. In 2005 he became a Fellow of the Frankfurt Institute for Advanced Studies (FIAS), in Frankfurt am Main, Germany. In 2006 he received a Marie Curie Excellence Center Award of the European Union. Since 2007 he is the Johanna Quandt Research Professor for Theoretical Life Sciences at FIAS. He also holds professorships at the Department of Physics and the Department of Computer Science and Mathematics at the Goethe University in Frankfurt am Main, Germany. In 2019 he obtained a visiting professorship at the Université Clermont-Auvergne, France. His research interests span Computational Neuroscience, Machine Learning, and Developmental Robotics.

MULTIPLE INSTANCE LEARNING FOR BRAIN TUMOR DETECTION FROM MAGNETIC RESONANCE SPECTROSCOPY DATA

A PREPRINT

Diyuan Lu[¶*] **Gerhard Kurz**[*†] **Nenad Polomac**[‡] **Iskra Gacheva**[‡] **Elke Hattingen**[‡] **Jochen Triesch**[¶]

ABSTRACT

We apply deep learning (DL) on Magnetic resonance spectroscopy (MRS) data for the task of brain tumor detection. Medical applications often suffer from data scarcity and corruption by noise. Both of these problems are prominent in our data set. Furthermore, a varying number of spectra are available for the different patients. We address these issues by considering the task as a multiple instance learning (MIL) problem. Specifically, we aggregate multiple spectra from the same patient into a "bag" for classification and apply data augmentation techniques. To achieve the permutation invariance during the process of bagging, we proposed two approaches: (1) to apply min-, max-, and average-pooling on the features of all samples in one bag and (2) to apply an attention mechanism. We tested these two approaches on multiple neural network architectures. We demonstrate that classification performance is significantly improved when training on multiple instances rather than single spectra. We propose a simple oversampling data augmentation method and show that it could further improve the performance. Finally, we demonstrate that our proposed model outperforms manual classification by neuroradiologists according to most performance metrics.

Keywords Tumor detection · Multiple instance learning · Machine learning · Magnetic resonance spectroscopy (MRS)

1 Introduction

We study the problem of brain tumor detection from MRS data. A brain tumor is the abnormal growth of the brain tissue, which can be benign or cancerous. In clinical practice, MRS is a common non-invasive tool used to identify a brain tumor, because it can be easily acquired alongside commonplace MR imaging procedures and it uniquely reflects the biochemical composition of the brain tissue *in situ*. MRS measures the resonant frequency shift of a chemically bound hydrogen atom (i.e., a proton), which characterizes different physiological or pathological brain metabolites. There has been increasing interest in MRS for clinical use because of the semiautomatic data acquisition, processing, and quantification [Ranjith et al., 2015, Hatami et al., 2018, González-Navarro and Belanche-Muñoz, 2009, Olliverre et al., 2018, Cruz-Barbosa and Vellido, 2011]. However, the interpretation of MRS spectra is traditionally performed by human radiologists based on the concentration ratios of certain metabolites. In contrast, we train a model to learn informative features from the spectra as a whole.

A common problem with MRS data is that they are often corrupted by noise from head movements during the procedure or baseline distortions of the spectrum. This poses difficulties in the MRS data interpretation. Additionally, labels are only provided per patient and not per voxel, which could introduce labeling noise as spectra from the tumor-affected hemisphere can be falsely labeled as "tumor" even though they contain healthy brain tissue.

Our contributions are summarized as follows.

[*]Both authors contribute equally to the paper. [¶]D. Lu and J. Triesch are with Frankfurt Institute for Advanced Studies, 60438 Frankfurt am Main, Germany. e-mail: elu, triesch@fias.uni-frankfurt.de.

[†]G. Kurz is an independent researcher. e-mail: kurz.gerhard@gmail.com

[‡]N. Polomac, I. Gacheva and E. Hattingen are with the Institute for Neuroradiology at Frankfurt university hospital, 60528 Frankfurt am Main, Germany. e-mail:Nenad.Polomac@kgu.de, elke.hattingen@kgu.de

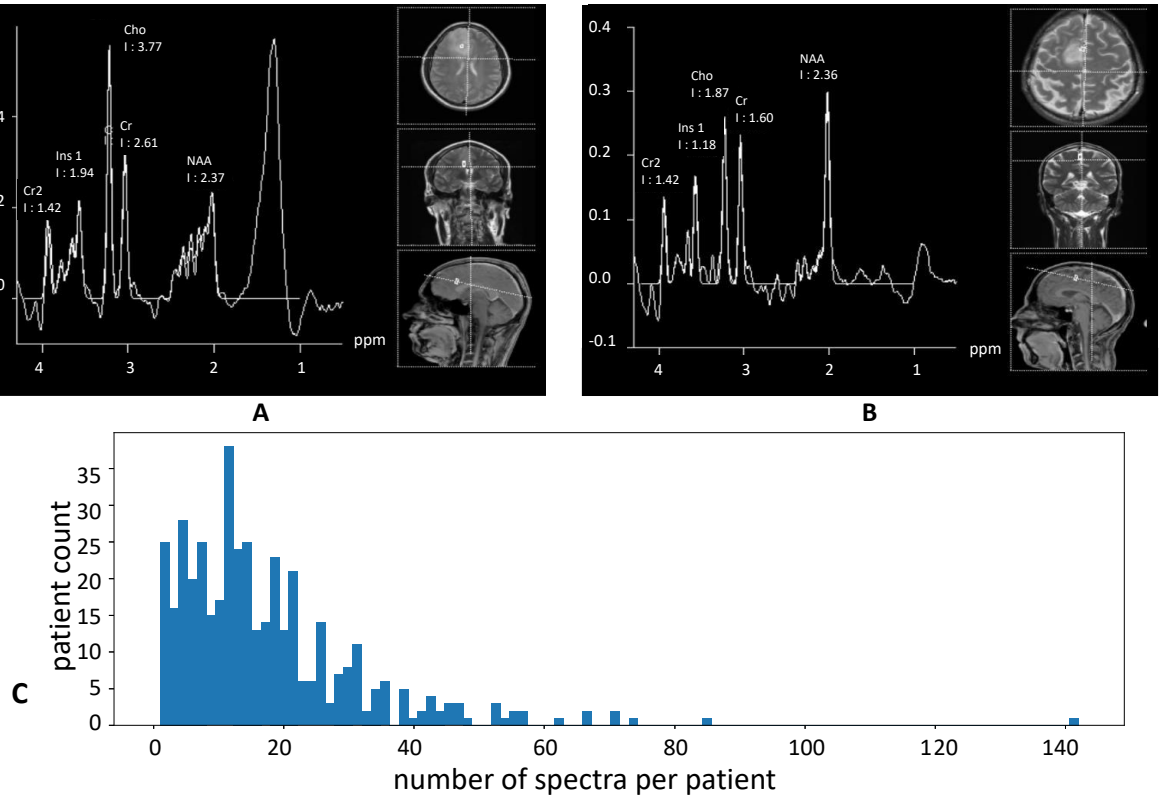


Figure 1: Overview of the MRS data used in this paper. Each spectrum is a data array with 288 data points with the x-axis indicating the position of different metabolites and the y-axis indicating the intensity of the corresponding metabolites. Several spectra may stem from the same patient. **A.** An example of a **tumor** MR spectrum. **B.** An example of a **non-tumor** MR spectrum. **C.** Histogram of the number of spectra per patient (17 ± 15, mean $\pm$ standard deviation).

- We present a multiple-instance-learning (MIL)-based framework for MRS-based tumor detection that performs patient-wise classification.
- We propose two modules to achieve permutation invariance when processing bags of instances simultaneously, i.e., an attention module and the concatenation of max-, min-, and average-pooling, which we refer to as the "3Pool" module.
- We demonstrate that our proposed modules can be easily plugged in any given DNN-based model and improve the classification performance.
- We evaluate the proposed method with a leave-patient-out cross validation scheme, which carefully tests the trained model on data from unseen patients. We also show that our method is even able to outperform human neuroradiologists.

2 Related Work

Modern machine learning approaches based on deep neural networks (DNNs) have recently obtained impressive results in a range of classification tasks, sometimes even outperforming human experts. These successes are based on, amongst others, (1) better learning algorithms, 2) fast computing hardware, and 3) large, carefully annotated data sets. This has motivated a range of applications in oncology such as tumor detection, tumor segmentation, tumor progression estimationLin et al. [2019], Capper et al. [2018], Park et al. [2019], Pereira et al. [2016], Ranjith et al. [2015], tumor grade classification [Ranjith et al., 2015], etc. However, acquiring the required labeled data is often hard to achieve or expensive in certain medical applications where the numbers of patients may be quite small. Multiple instance learning (MIL) is a framework to handle scenarios where detailed annotations for each individual instance is noisy, laborious to obtain, or simply not available. It tries to make a decision based on a set of single instances instead of a decision for

each single instance. MIL has been widely used in medical applications such as breast cancer detection [Sudharshan et al., 2019, Conjeti et al., 2017, Sadafi et al., 2020] and other forms of computer assisted diagnosis [Fung et al., 2007, Liu et al., 2018].

Applying machine learning methods to medical applications with MRS data is gaining more and more momentum, for example in brain tumor detection [González-Navarro and Belanche-Muñoz, 2009, Cruz-Barbosa and Vellido, 2011, Rao et al., 2015], brain tumor segmentation [Dvořák and Menze, 2015, Pereira et al., 2016], breast tumor detection [Tavolara et al., 2019, Ren et al., 2015], and tumor motion prediction [Lin et al., 2019]. There is also work to investigate the effect of the length of the echo time used to perform MR spectroscopy for the tumor detection [González-Navarro and Belanche-Muñoz, 2009]. Olliverre et al. [2018] proposed to use generative-adversarial-network-based model to synthesize MRS data with real-world appearance and features for deep model training. Cruz-Barbosa and Vellido [2011] proposed a variant of generative topographic mapping method for diagnostic discrimination between different brain tumor pathologies and the outcome prediction.

Noisy labels are ubiquitous in the real world. We use the term *noisy labeling* to refer to annotations that are incorrect, i.e., due to the labeling procedure, the label assigned patient-wise, so they reflect the overall diagnosis rather than properties of a specific spectrum. Noisy labels are posing a non-trivial problem in deep model learning when an increasing ability to fit noise is accompanied with deeper layers. Given the ubiquity and importance of coping with noisy labeling, many works have been devoted to combating this problem. Some of them start with a small set of clean expert-labeled data [Han et al., 2018, Li et al., 2017, Veit et al., 2017, Albarqouni et al., 2016], but this may not be trivial to obtain. Consequently, models that can learn directly with noisy labels [Han et al., 2018, Smyth et al., 2019, Rolnick et al., 2017] are highly desirable.

Multiple instance learning (MIL) is a framework to combat the problems, where detailed annotation for each single instance is noisy, or is laborious to obtain, or simply not available. Single-Instance Learning is a "naive" approach that assigns all instances in one bag the same label as its bag, which might lead to mislabeling negative instances in positive bags [Ray and Craven, 2005]. Andrews et al. [2002] proposed to modify the standard SVM so that the MI assumption that at least one instance in each bag is positive is applicable. The normalized set kernel (NSK) and statistics kernel methods apply kernels to map the whole bags of instances into features, then use the standard SVM to make the classification on the bag level Gärtner et al. [2002]. MIL has also been widely used in medical applications such as breast cancer detection [Sudharshan et al., 2019, Conjeti et al., 2017], computer assist diagnosis [Fung et al., 2007], brain disease diagnosis [Liu et al., 2018], lung cancer diagnosis [Ozdemir et al., 2019], blood cell disorder analysis [Sadafi et al., 2020], etc.

3 Methods

3.1 Data

In this study, We use 1H-MR-spectroscopy data collected from 435 patients recorded in the Institute for Neuroradiology of the University Hospital in Frankfurt between 01/2009 to 3/2019. They were reviewed retrospectively and have been completely anonymized for this study. The patients were suffering from either glial or glioneuronal first diagnosed tumors (the *tumor* group, 266 patients) or other non-neoplastic lesions, e.g., demyelination, gliosis, focal cortical dysplasia, enlarged Virchow-Robin spaces or similar (the *non-tumor* group, 156 patients). The tumor group included all spectra from the tumor-affected hemisphere. The non-tumor group consisted of spectra from both hemispheres of the patients.

As a result, 7442 spectra (3388 non-tumor and 4054 tumor) were selected for further analysis. The obtained MRS examples are saved as 1-d arrays with 288 data points, i.e., in shape (288×1), shown in Fig. 1A, B, where the y-axis shows signal intensities of different metabolites, and the x-axis represents the chemical shift positions in ppm indicating various metabolites. The indices correspond to the position of metabolites and the values indicate signal intensities of corresponding metabolites. We normalize each spectrum to zero mean and unit variance. All spectra from the same patient are labeled with the patient's diagnosis, i.e., all spectra from one tumor patient will be labeled as *tumor*, and all spectra from one non-tumor patient would be labeled as *non-tumor*.

There is a huge variance in the number of spectra per patent in our data set - some patients have dozens of spectra and some have just a few spectra or even just a single one. A histogram of the number of spectra per patient is shown in Fig. 1C. Due to the fairly limited number of patients, machine learning methods trained on this data set are prone to overfitting, therefore applying out-of-the-box methods would not yield satisfactory results. Each spectrum describes the biochemical composition of one voxel of brain tissue. We propose to perform classification not on a single spectrum, but on a bag of spectra from this patient. Specifically, we create bags of spectra from each patient for training and validation.

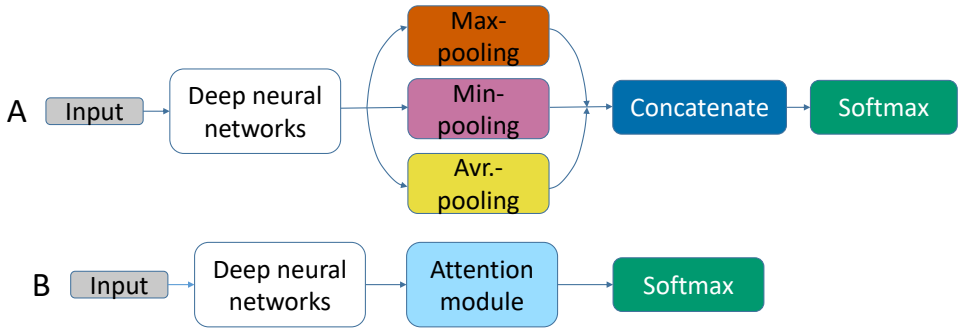


Figure 2: Overview of the proposed framework with two proposed permutation invariant modules, which can be plugged in any DNN-based models. **A.** The proposed "3Pool" module. **B.** The proposed attention module.

3.2 Patient-wise Data Preparation

We have MRS spectra from a total number of P patients, the total number of spectra for patient p is N_p. We generate bags of spectra consisting of a fixed number $\in \mathbb{N}$ of spectra from each patient by sampling from all spectra of the patient with replacement during training. Each bag is in shape $\times 288$. The bags from patient p are denoted as $\mathbf{X}^p = \{\mathbf{x}_1^p, \ldots, \mathbf{x}_{p_b}^p\}$, where p_b is the total number of bags generated for patient p. Since, this is a combinatorial problem, we could potentially generate millions of samples. This could be viewed as a data augmentation (DA) process. However, the more bags we generate from one patient, the less diversity we introduce through the DA and the worse the network is at generalization. Empirically, we set the number of generated bags of one patient to three times their single spectra count. Of course, further exploration of the optimum number of spectra to use might be beneficial in the future. Each training bag is provided a class label $y^p \in \{\textit{tumor}, \textit{non-tumor}\}$ based on the diagnosis of the patient. More formally, our goal is to learn a function f, which takes a set of spectra $\mathbf{x}^p = \{x_i^p, \ldots, x_N^p\}$ from patient p, and output the classification decision $\hat{y}^p$. The function f processes all spectra at the same time and generates a final predicted label $\hat{y}$. The training objective is the classic cross entropy loss:

$$\theta \min \mathbb{E}_{P(\mathbf{x},\hat{y})}[-\log P_\theta(y = \hat{y}|\mathbf{x})], \tag{1}$$

where θ refers to the parameters of the function f.

The ability of the classifier to generalize to new previously unseen patients is of great clinical importance. Therefore, we apply a 5-fold leave-subjects-out cross validation scheme. To be specific, we divide the patient list into 5 sub-lists, each with around 80 patients. In each cross validation set, we withhold the data from the patients of one sub-list, while we train and validate on the data from the other sub-lists. During training and validation, we adopt a 4:1 split ratio of all generated bags. During testing, we switch off the data augmentation strategy and only allow the minimal repetition of the spectra to fill up the last bag, which may be only partially filled otherwise. This makes sure that the number of bags to generate for patient p follows

$$p_b = \{\, 1 \,, if N_p \leq, N_p/, if N_p > \,. \tag{2}$$

3.3 Network Structure

When working with bags of MRS spectra, we note that the order of the stacked spectra was randomly chosen and should not affect the result of the network. Being invariant to the order of the spectra can either be achieved by augmenting with shuffled data, which is an approximation, or by designing the network architecture in such a way that the output of the network is independent of the order of the spectra in the input. In this work, we compared both approaches. For the former, we have described the data augmentation that we use to generate bags of training samples in section 3.2. For the latter, we proposed two modules that can be easily plugged in any DNN-based models: (1) to aggregate the minimum-, maximum- and mean-pooling of the feature maps which yields exact order invariance, (2) to leverage attention mechanism [Ilse et al., 2018, Sadafi et al., 2020], where different instances in the bag are assigned with different attention weights, which can be learned by the neural network. The schematic of propose method is shown in Fig. 2. The final extracted feature is a weighted average of features from all the instances in one bag. Since the attention weights depend on the instance itself and not the order, we can also achieve exact permutation invariance.

In this work, we test the two proposed modules on several network structures, i.e., a multi-layer perceptron (MLP), an Inception-variant tailored to MRS data, and a CNN model inspired by Hatami *et al.* Hatami et al. [2018]. An Inception model is a successful neural network structure proposed to scale up convolution networks in efficient ways Szegedy et al. [2016]. In our implementation, we only preserve the first five inception blocks from the original InceptionV3

model Szegedy et al. [2016] and reduce the number of filters in each block compared to the original configuration due to a lower complexity of our MRS data compared to the image data. In the MLP model, there are three dense layers with 128, 32, and 2 dense units, respectively, as shown in Fig. 2B. In the model inspired from Hatami *et al.*, we omit the last convolutional layer with 512 kernels and the max-pooling layer, since the length of our data is smaller than theirs. Furthermore, for each model, we consider two variants, i.e., the one with concatenated max-, min-, and average-pooling, denoted "3Pool" and the other with an attention module, denoted "Att". Note that the feature extraction in each dense layer is performed on the single instance level, i.e., the convolution is only done horizontally with the kernel height as one. The feature maps are then either pooled and concatenated in a "3Pool" branch, or processed by the attention module.

3.4 Attention Module

In order to weigh the different samples contained in a bag, we make use of the attention mechanism proposed by Ilse et al. [2018]. The idea is to introduce a layer whose output z is a weighted average $\mathrm{z} = \sum_{k=1}^{a}{}_{k} h_k$ of the inputs h_k with weights $\mathrm{a}_k = \frac{\exp\left(w^T \tanh(V h_k^T)\right)}{\sum_{k=1}^{\exp\left(w^T \tanh(V h_k^T)\right)}}$ where $w \in \mathbb{R}^{1 \times N_{att}}$ and $V \in \mathbb{R}^{N_{att} \times L_{h_k}}$ are learned parameters of the layer. N_{att} is the number of attention heads and L_{h_k} is the dimension of the hidden feature h_k. As each a_k depends on the values inside h_k, the weights are different in each bag and can take the concrete values inside the input bag into account. Note that the output z is independent of the order of the inputs h_k

3.5 Training Procedure

The network is trained with randomly initialized weights using the Adam optimizer with default parameters and a mini-batch size of 32. The model is trained on a Windows machine with an Intel(R) Core i7-4770 CPU, 16 GB RAM and a GeForce GTX1060 GPU with 6GB of memory. The training and takes less than 3 minutes for 30 training epochs.

4 Results

4.1 Overall Performance with Ablation

To evaluate performance, we use the area under the receiver operating characteristic (ROC) curve, the F1-Score and the Matthews correlation coefficient (MCC). The ROC curve is constructed by varying the classification threshold and calculating the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) rates. We report classification accuracy, area under the ROC curve (AUC), F1-score = $\frac{2TP}{2TP+FP+FN}$, and MCC = $\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$. The MCC is generally considered as a balanced measure which takes into account TP, TN, FP, and FN, and it can be used even if the classes are not balanced. We also conducted ablation studies on the effectiveness of data augmentation on different network structures. Moreover, we compared our method to three other baseline methods, i.e., the support vector machine approaches by Ray-MISVM Ray and Craven [2005], MI-SVM Andrews et al. [2002], and NSK Gärtner et al. [2002]. For this purpose, we used the implementation from Doran [2019].

Empirically, we found that using 31 spectra per bag yields relatively good results. Therefore, we report the averaged performance metrics with the default $= 31$ across all cross validation sets. The results averaged across all leave-subjects-out cross validation sets are shown in Table 1. In addition to the comparison on multiple instances learning, we also ran all the models (1) with single instances, denoted with "(SI)", (2) with the oversampling data augmentation, denoted with " + DA". From Table 1, we made the following observations and possible explanations. Firstly, the CNN network inspired from Hatami *et al.* with the proposed "3Pool" module achieved the best results: a bag AUC of 0.82, a patient-wise AUC of 0.82, an F1-score of 0.78 and an MCC of 0.46. Secondly, when grouping multiple instances into bags for training without any data augmentation, models with low complexity, indicated by the number of trainable parameters, show a performance deterioration and models with a large number of trainable parameters still show an improvement in the performance. One contributing factor might be that the number of total training samples are significantly reduced when changing from the SI learning case to MI learning, thus the generalization ability is not fully explored. Thirdly, the "3Pool" module works the best with high complexity models such as Hatami-model and Inception. Thirdly, data augmentation (" + DA") almost always helps improve the performance, except in the case of MI-SVM. Thirdly, of the two proposed approaches to achieve permutation invariance, i.e., (1) using max-, min-, and average-pooling of feature maps before the softmax activation, and (2) the attention-weighted average of feature maps before the softmax activation, we found that the first approach works better when combined with the Inception network, but the second approach is superior when using the MLP. Thus, neither approach is clearly superior and the choice of method needs to be made depending on the particular structure of the underlying neural network.

Table 1: Performance matrices averaged across **five-fold cross validation data sets** of proposed method compared to other baseline methods. The results are shown in **mean ± standard deviation**. MCC: Matthews correlation coefficient, AUC: area under ROC curve. SI: single instance. Baseline MIL models: MI-SVM [Andrews et al., 2002], Ray-MISVM [Ray and Craven, 2005], and NSK [Gärtner et al., 2002]. MLP: multi-layer perceptron.

	Bag AUC	**Patient AUC**	**F1-score**	**MCC**	**# Trainables**
Ray-MISVM Ray and Craven [2005] (SI)	0.73 ± 0.07	0.74 ± 0.06	0.69 ± 0.06	0.31 ± 0.12	$\sim$600
Ray-MISVM Ray and Craven [2005]	0.63 ± 0.02	0.59 ± 0.02	0.54 ± 0.03	0.19 ± 0.05	$\sim$600
Ray-MISVM Ray and Craven [2005] + DA	0.73 ± 0.09	0.73 ± 0.08	0.71 ± 0.10	0.35 ± 0.18	$\sim$600
MI-SVM Andrews et al. [2002] (SI)	0.71 ± 0.04	0.74 ± 0.06	0.68 ± 0.04	0.30 ± 0.09	$\sim$600
MI-SVMAndrews et al. [2002]	0.69 ± 0.07	0.69 ± 0.07	0.69 ± 0.07	0.30 ± 0.12	$\sim$600
MI-SVMAndrews et al. [2002] + DA	0.69 ± 0.08	0.69 ± 0.07	0.70 ± 0.07	0.30 ± 0.12	$\sim$600
NSK Gärtner et al. [2002] (SI)	0.72 ± 0.05	0.72 ± 0.05	0.71 ± 0.05	0.34 ± 0.09	$\sim$600
NSK Gärtner et al. [2002]	0.70 ± 0.06	0.69 ± 0.06	0.69 ± 0.04	0.30 ± 0.11	$\sim$600
NSK Gärtner et al. [2002] + DA	0.74 ± 0.04	0.74 ± 0.04	0.72 ± 0.03	0.35 ± 0.06	$\sim$600
MLP (SI)	0.73 ± 0.04	0.77 ± 0.05	0.69 ± 0.05	0.30 ± 0.08	41,314
MLP-3Pool	0.68 ± 0.07	0.68 ± 0.08	0.69 ± 0.04	0.30 ± 0.09	41,314
MLP-3Pool + DA	0.72 ± 0.11	0.72 ± 0.10	0.70 ± 0.07	0.33 ± 0.15	41,314
MLP-Att	0.78 ± 0.08	0.78 ± 0.08	0.73 ± 0.08	0.37 ± 0.17	41,220
MLP-Att + DA	0.79 ± 0.06	0.79 ± 0.05	0.74 ± 0.05	0.42 ± 0.11	41,220
Hatami (SI)	0.67 ± 0.03	0.72 ± 0.04	0.65 ± 0.03	0.23 ± 0.06	488,514
Hatami-3Pool	0.77 ± 0.07	0.76 ± 0.06	0.72 ± 0.05	0.36 ± 0.11	488,514
Hatami-3Pool + DA	$\mathbf{0.82 \pm 0.07}$	$\mathbf{0.82 \pm 0.06}$	$\mathbf{0.78 \pm 0.08}$	$\mathbf{0.46 \pm 0.19}$	488,514
Hatami-Att	0.80 ± 0.05	0.80 ± 0.04	0.73 ± 0.05	0.37 ± 0.14	507,012
Hatami-Att + DA	0.81 ± 0.06	0.81 ± 0.05	0.75 ± 0.08	0.43 ± 0.16	507,012
Inception-3Pool (SI)	0.73 ± 0.07	0.76 ± 0.07	0.69 ± 0.07	0.32 ± 0.14	345,098
Inception-3Pool	0.77 ± 0.07	0.76 ± 0.06	0.72 ± 0.05	0.36 ± 0.11	345,098
Inception-3Pool + DA	0.79 ± 0.05	0.79 ± 0.05	0.74 ± 0.05	0.39 ± 0.09	345,098
Inception-Att	0.75 ± 0.06	0.75 ± 0.05	0.72 ± 0.05	0.36 ± 0.10	345,116
Inception-Att + DA	0.76 ± 0.07	0.75 ± 0.07	0.73 ± 0.05	0.38 ± 0.11	345,116

4.2 Human vs. Machine

We compared the performance of implemented DNN models to that of human neuroradiologists on one randomly selected test set, which has 844 spectra from around 42 patients. The result is shown in Tab. 2. For the collection of the classification results of neuroradiologists, we divided the test set into eight subsets and each subset was assigned to one of eight neuroradiologists. The neuroradiologists' performance therefore represents the collective effort of eight individuals, which is faithfully reflect the clinical practice. The data shows that the performance of our proposed method is better on almost all performance metrics except the MCC. The reason is that the neuroradiologists achieved a specificity of 0.88 but at a cost of a low sensitivity of 0.54. This may reflect that neuroradiologists assign different "costs" to false positive vs. false negative classifications.

4.3 Attention Visualization

Further, we show two bags of samples from each class with color-coded attention during testing, shown in Fig 3. We can see that features of spectra in one bag are very heterogeneous exhibiting different peak ratios, peak positions, etc. Note that, samples with high attention might be stereotypical of that class or raising a red flag for that class decision. One benefit of visualizing the attention assignment is that it provides not only a final classification result but also the contextual information of the same patient's brain tissue. This could provide more information for the MRS data interpretation. The common metabolites from left to right in our data are creatine2 (Cr2, 3.9 ppm), myo-inositol and glycine (MI/Gly, $\sim$ 3.5 ppm), Myo-inositol (Ins, 3.61 ppm), choline (Cho, 3.19 ppm), creatine (Cr, 3.03 ppm), Glutamin (Glu, 2.2 – 2.4 ppm), N-acetyl aspartate (NAA, 2.01 ppm), lactate (Lac, 1.4 ppm), and Lipids (Lip, 0.9 ppm) Faghihi et al. [2017], Fan et al., Fan [2006], Rae, Hattingen et al. [2009]. There are several indicative features in MRS data that

Table 2: Performance on **withheld neuroradiologist-labeled data set** of all models. SIC: single-instance classification. MCC: Matthews correlation coefficient, AUC: area under ROC curve. SI: single instance. Baseline MIL models: MI-SVM [Andrews et al., 2002], Ray-MISVM [Ray and Craven, 2005], and NSK [Gärtner et al., 2002]

	Bag AUC	**Patient AUC**	**F1-score**	**MCC**
Neuroradiologists	–	–	0.56	**0.58**
Ray-MISVM Ray and Craven [2005] (SI)	0.64 ± 0.04	0.60 ± 0.03	0.52 ± 0.03	0.14 ± 0.06
Ray-MISVM Ray and Craven [2005]	0.62 ± 0.02	0.59 ± 0.02	0.55 ± 0.04	0.12 ± 0.08
Ray-MISVM Ray and Craven [2005] + DA	0.63 ± 0.02	0.59 ± 0.02	0.55 ± 0.05	0.16 ± 0.10
MI-SVM Andrews et al. [2002] (SI)	0.67 ± 0.02	0.65 ± 0.04	0.58 ± 0.05	0.29 ± 0.08
MI-SVM Andrews et al. [2002]	0.63 ± 0.03	0.59 ± 0.03	0.60 ± 0.04	0.26 ± 0.09
MI-SVM Andrews et al. [2002] + DA	0.65 ± 0.02	0.60 ± 0.03	0.62 ± 0.03	0.26 ± 0.05
NSK Gärtner et al. [2002] (SI)	0.70 ± 0.02	0.68 ± 0.03	0.58 ± 0.02	0.27 ± 0.03
NSK Gärtner et al. [2002]	0.69 ± 0.05	0.65 ± 0.06	0.60 ± 0.05	0.23 ± 0.09
NSKGärtner et al. [2002] + DA	0.73 ± 0.05	0.69 ± 0.05	0.66 ± 0.06	0.35 ± 0.10
MLP-3Pool (SI)	0.74 ± 0.06	0.74 ± 0.06	0.61 ± 0.07	0.32 ± 0.12
MLP-3Pool	0.77 ± 0.04	0.70 ± 0.05	0.68 ± 0.05	0.38 ± 0.11
MLP-3Pool + DA	0.75 ± 0.05	0.69 ± 0.06	0.67 ± 0.03	0.35 ± 0.08
MLP-Att	0.76 ± 0.04	0.70 ± 0.04	0.65 ± 0.02	0.33 ± 0.04
MLP-Att + DA	0.78 ± 0.03	0.72 ± 0.03	0.65 ± 0.02	0.33 ± 0.06
Hatami (SI)	0.66 ± 0.02	0.69 ± 0.04	0.53 ± 0.02	0.22 ± 0.03
Hatami-3Pool	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.80 \pm 0.03}$	0.74 ± 0.03	0.49 ± 0.06
Hatami-3Pool + DA	0.84 ± 0.02	0.78 ± 0.03	0.70 ± 0.01	0.43 ± 0.03
Hatami-Att	0.83 ± 0.04	0.78 ± 0.04	0.71 ± 0.03	0.45 ± 0.05
Hatami-Att + DA	0.85 ± 0.02	0.80 ± 0.03	0.74 ± 0.02	0.49 ± 0.03
Inception-3Pool (SI)	0.73 ± 0.06	0.70 ± 0.05	0.61 ± 0.06	0.34 ± 0.10
Inception-3Pool	0.83 ± 0.04	0.77 ± 0.04	$\mathbf{0.75 \pm 0.02}$	0.56 ± 0.05
Inception-3Pool + DA	0.81 ± 0.04	0.74 ± 0.04	0.70 ± 0.05	0.43 ± 0.10
Inception-Att	0.82 ± 0.04	0.76 ± 0.04	0.74 ± 0.06	0.50 ± 0.11
Inception-Att + DA	0.79 ± 0.03	0.74 ± 0.04	0.70 ± 0.03	0.43 ± 0.05

are clinically relevant. For example, in tumor spectra, there are weakened Cr and Ins Faghihi et al. [2017], reduced NAA concentration Faghihi et al. [2017], elevated Cho, Glu, Lac, Lip peaks Rae, Fan [2006], Faghihi et al. [2017], elevated MI/Gly Hattingen et al. [2009].

In Fig. 3, we can see that in the *non-tumor* group, the high attention weights are assigned to samples with flat Lip, flat Lac Rae, high and narrow NAA (low 2.0 – 2.5 ppm), clear Cr/Cho ratio > 1, etc. For the *tumor* group, the high attention weights are often assigned to instances with low NAA with elevated Glu, high Lac, high Lip, clear Cr/Cho ratio < 1, as shown in Rae, Fan [2006], Hattingen et al. [2009], Faghihi et al. [2017]

4.4 Varying the Bag Size

To investigate the effect of the number of samples per bag, we vary the value from one (corresponding to single instance classification) to 51. The AUC as a function of the number is shown in Fig 4. From this experiment, we made the following observations. Firstly, for all models, learning from the bags of multiple instances is better than learning from a single instance. The performance is significantly improved when increases from one to six, and then this improvement attenuated after $= 6$ in all models. Secondly, the performance with the attention module did not show a deterioration with an increasing in all models. However, in the MLP model, the performance degraded after $= 6$ with the "3Pool" module.

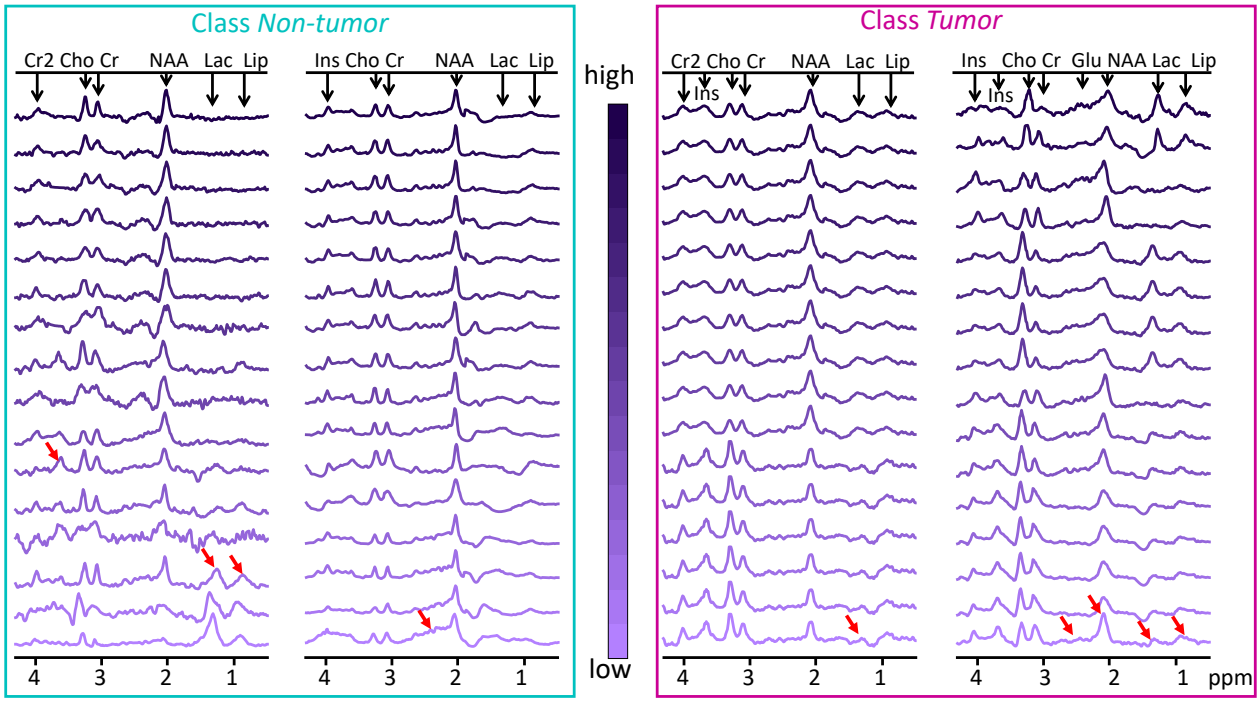


Figure 3: Exemplar bags (column) of MRS spectra with attention, color-coded with shades in a descending order. Red arrows in instances with relatively low attention show the features from the opposite class.

5 Conclusion

This paper presents a novel framework for tumor detection based on multiple instance (MI) learning with noisily-labeled MRS data. We proposed two modules to achieve permutation-invariance within each bag: (1) an attention module and (2) a "3Pool" module with max-, min-, and average-pooling. Moreover, we applied data augmentation to generate bags of instances from each patient, which expanded the total training data size as well as increased the variance in the training data. We applied these two modules on several popular DNN models, i.e., an MLP, an Inception-variant, and a CNN-based model inspired by Hatami et al. Hatami et al. [2018]. We conducted a thorough comparison between the different models as well as three conventional SVM-based MI methods. We also carried out an ablation study regarding the effect of the data augmentation for all models. We observed the MI SVMs do not perform well on our data set. The data augmentation almost always improved the performance compared to the counterpart without augmentation, except in the case of Andrews et al. [2002]. In the Hatami-model and the Inception model, the proposed "3Pool" module achieved slightly better performance than the "Att" module. However, in the MLP model, the proposed "Att" module was superior. The best results of all experimented configurations were obtained by the Hatami-model with the proposed "3Pool" module and data augmentation: a bag AUC of 0.82, a patient-wise AUC of 0.82, an F1-score of 0.78, and an MCC of 0.46. We showed that our MI-based approach significantly improved the performance compared to single instance classification (t-test with a p-value of ≤ 0.004) and that applying data augmentation for generating more training data is beneficial to obtain good results, however it does not rise to the level of being statistically significant. We also demonstrate that the proposed method outperforms human radiologists in terms of F1-score while achieving a similar MCC. The limitation of this work is that the results are obtained from a data set collected from a single site. Due to the factors such as the variability of data acquisition procedures, the diverse patient populations, the generalization ability of the proposed method to other MRS data sets is not demonstrated. Furthermore, so far we only experimented with a very simple data augmentation method. Further exploration of other data augmentation strategies such as mixup, adding noise, scaling amplitude, etc., might be interesting in the future. A further inspection of the different effects of "Att" and "3Pool" to the learning of different networks is also of interest. So far, we used a stratified sampling strategy, i.e., the more single spectra one patient has, the more bags we generate. This could potentially introduce bias. In the future, we could fix the number of bags to generate for all patients to eliminate the bias introduced by the current method. Furthermore, we could explore other statistics within the bag such as the median and the interquartile range. Adding explainable machine learning methods is also beneficial for promoting the approach for clinical practice.

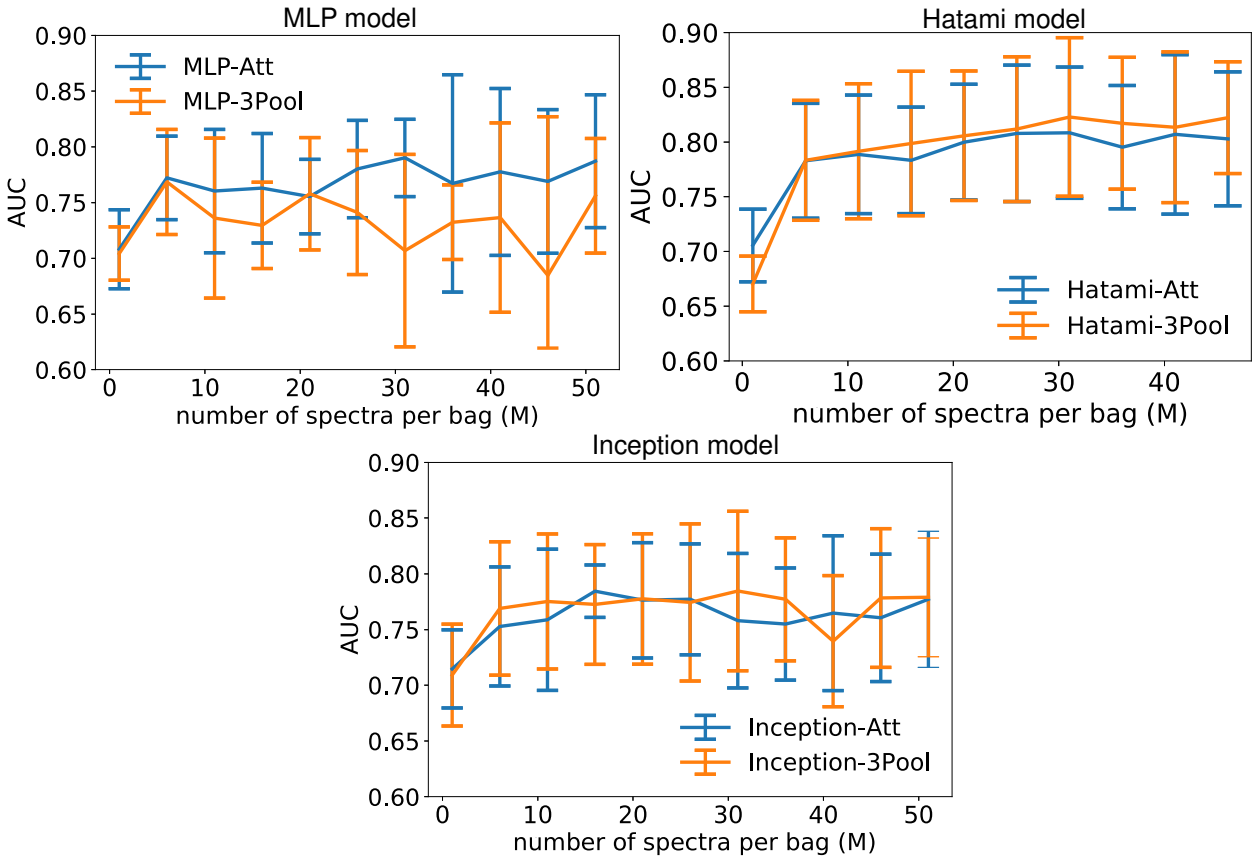


Figure 4: Averaged ROC-AUC as a function of the number of instances per bag across five leave-patients-out cross validation sets for our proposed methods. The errorbars represent one standard deviation.

Finally, we would like to investigate the behaviour of the proposed approaches on further data sets collected at other sites.

References

S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, 35(5):1313–1321, 2016.

S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, volume 2, pages 561–568. Citeseer, 2002.

D. Capper, D. T. Jones, M. Sill, V. Hovestadt, D. Schrimpf, D. Sturm, C. Koelsche, F. Sahm, L. Chavez, D. E. Reuss, et al. Dna methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469, 2018.

S. Conjeti, M. Paschali, A. Katouzian, and N. Navab. Deep multiple instance hashing for scalable medical image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 550–558. Springer, 2017.

R. Cruz-Barbosa and A. Vellido. Semi-supervised analysis of human brain tumours from partially labeled mrs information, using manifold learning models. *International journal of neural systems*, 21(01):17–29, 2011.

G. Doran. Misvm: Multiple-instance support vector machines, 2019. URL `https://github.com/garydoranjr/misvm`.

P. Dvořák and B. Menze. Local structure prediction with convolutional neural networks for multimodal brain tumor segmentation. In *International MICCAI workshop on medical computer vision*, pages 59–71. Springer, 2015.

R. Faghihi, B. Zeinali-Rafsanjani, M.-A. Mosleh-Shirazi, M. Saeedi-Moghadam, M. Lotfi, R. Jalli, and V. Iravani. Magnetic resonance spectroscopy and its clinical applications: a review. *Journal of medical imaging and radiation sciences*, 48(3):233–253, 2017.

G. Fan. Magnetic resonance spectroscopy and gliomas. *Cancer Imaging*, 6(1):113–115, 2006.

G. Fan, B. Sun, Z. Wu, Q. Guo, and Y. Guo. In vivo single-voxel proton mr spectroscopy in the differentiation of high-grade gliomas and solitary metastases. 59(1):0–85.

G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao. Multiple instance learning for computer aided diagnosis. *Advances in neural information processing systems*, 19:425, 2007.

T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *ICML*, volume 2, page 7, 2002.

F. F. González-Navarro and L. A. Belanche-Muñoz. Using machine learning techniques to explore 1h-mrs data of brain tumors. In *2009 Eighth Mexican International Conference on Artificial Intelligence*, pages 134–139. IEEE, 2009.

B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, 2018.

N. Hatami, M. Sdika, and H. Ratiney. Magnetic resonance spectroscopy quantification using deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 467–475. Springer, 2018.

E. Hattingen, H. Lanfermann, J. Quick, K. Franz, F. E. Zanella, and U. Pilatus. 1 h mr spectroscopic imaging with short and long echo time to discriminate glycine in glial tumours. *Magnetic Resonance materials in physics, biology and medicine*, 22(1):33, 2009.

M. Ilse, J. M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. *CoRR*, abs/1802.04712, 2018. URL `http://arxiv.org/abs/1802.04712`.

Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L. J. Li. Learning from Noisy Labels with Distillation. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:1928–1936, 2017.

H. Lin, W. Zou, T. Li, S. J. Feigenberg, B.-K. K. Teo, and L. Dong. A super-learner model for tumor motion prediction and management in radiation therapy: Development and feasibility evaluation. *Scientific reports*, 9(1):1–11, 2019.

M. Liu, J. Zhang, E. Adeli, and D. Shen. Landmark-based deep multi-instance learning for brain disease diagnosis. *Medical image analysis*, 43:157–168, 2018.

N. Olliverre, G. Yang, G. Slabaugh, C. C. Reyes-Aldasoro, and E. Alonso. Generating magnetic resonance spectroscopy imaging data of brain tumours from linear, non-linear and deep learning models. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 130–138. Springer, 2018.

O. Ozdemir, R. L. Russell, and A. A. Berlin. A 3d probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose ct scans. *IEEE transactions on medical imaging*, 39(5):1419–1429, 2019.

E. K. Park, K.-s. Lee, B. K. Seo, K. R. Cho, O. H. Woo, G. S. Son, H. Y. Lee, and Y. W. Chang. Machine learning approaches to radiogenomics of breast cancer using low-dose perfusion computed tomography: Predicting prognostic biomarkers and molecular subtypes. *Scientific reports*, 9(1):1–11, 2019.

S. Pereira, A. Pinto, V. Alves, and C. A. Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

C. Rae. Re: Magnetic resonance spectroscopy of the brain: review of metabolites and clinical applications. *Clinical Radiology*, 64(10):0–1043.

G. Ranjith, R. Parvathy, V. Vikas, K. Chandrasekharan, and S. Nair. Machine learning methods for the classification of gliomas: Initial results using features extracted from mr spectroscopy. *The neuroradiology journal*, 28(2):106–111, 2015.

V. Rao, M. S. Sarabi, and A. Jaiswal. Brain tumor segmentation with deep learning. *MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS)*, 59, 2015.

S. Ray and M. Craven. Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the 22nd international conference on Machine learning*, pages 697–704, 2005.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. *CoRR*, abs/1705.10694, 2017.

A. Sadafi, A. Makhro, A. Bogdanova, N. Navab, T. Peng, S. Albarqouni, and C. Marr. Attention based multiple instance learning for classification of blood cell disorders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 246–256. Springer, 2020.

L. Smyth, D. Kangin, and N. Pugeault. Training-valuenet: Data driven label noise cleaning on weakly-supervised web images. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 307–312. IEEE, 2019.

P. Sudharshan, C. Petitjean, F. Spanhol, L. E. Oliveira, L. Heutte, and P. Honeine. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111, 2019.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

T. E. Tavolara, M. K. K. Niazi, V. Arole, W. Chen, W. Frankel, and M. N. Gurcan. A modular cgan classification framework: Application to colorectal tumor detection. *Scientific reports*, 9(1):1–8, 2019.

A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie. Learning from noisy large-scale datasets with minimal supervision. 2017.

Chapter 3

Conclusion and Outlook

In this thesis, I have addressed the problem of applying machine learning methods to medical applications in the context of epilepsy and brain tumor detection. In the early diagnosis of epilepsy project, we wished to understand the disease progression in the latent epileptogenesis phase and to make early predictions regarding whether the individual has a high risk of developing epilepsy. In the brain tumor detection project, we aimed to provide a pre-screening tool for tumor detection from magnetic resonance spectroscopy (MRS) data for clinical practice.

In tackling the aforementioned tasks, we have dealt with different medical data modalities, facing several ubiquitous problems in ML-based medical applications. We have also provided some solutions to these problems, which could be applied to other medical applications. The comparison between the two projects is summarized as follows.

3.1 Commonalities and Differences

Dataset sizes: In the early diagnosis of epilepsy with EEG, we collected a

vast amount of data from 10 rats. On average, we obtained one month's recordings, 24/7, from each rat. In contrast, in tumor detection, we collected over 7000 samples from over 400 patients and, on average, less than 20 samples per patient. Thus, in the second project, we were confronted with scarcity of data and, thus, had to employ different strategies of data augmentation, however, this was not required for the first project.

Leave-individual-out Cross-validation: In both projects, we were constrained either by the number of individuals in the experiment (in the epilepsy project), or by the numbers of samples from each individual (in the brain tumor detection project). Nevertheless, we need to ensure the generalization ability of the trained model to unseen individuals. This is of great importance especially in medical applications since DNN-based models have the tendency to overfit on the training data and struggle to obtain the same-level accuracy during testing as during the training. Ultimately, we want the pretrained DNN models to generalize well to new patients whose data is never accessible by the model. To ensure this generalization ability, we trained our models in a leave-individual-out cross-validation scheme, in which we trained the model with the data from a subset of individuals, and tested the model with unseen data from unseen individuals. To be specific, in the epilepsy project, in each cross-validation fold, we trained the model while completely withholding the data from one animal, and the final performance was averaged across all cross-validation folds. In the brain tumor detection project, we divided the total list of the patients into several sublists. For each sublist, we held out the data from the patients on the list for testing and trained with the data from other sublists. In this way, we can develop ML methods and assess the performance properly.

Label qualities: In the epilepsy project, the data was only annotated according to the timestamps when each rat underwent the epilepsy-inducing stimulation.

Based on these timestamps, we could generate "free", but relatively weak, labels such as "baseline" and "epileptogenesis". For tumor detection, the labels were obtained per patient through manual labeling, not per sample, however, since there may have been healthy tissue in the tumor patients and, likewise, tumor-like samples in the non-tumor patients. This, in turn, introduced label noise.

Individual variability: In the rodent epilepsy models, different rats responded differently to the same controlled stimulation and, thus, the damage induced by the stimulation varied. Therefore, the disease progression and the duration of epileptogenesis differed in the rodent models. In our preliminary experiment, we were able to train a simple classifier to identify individual animals based on the EEG traces. This meant that the identity information could also play a role in the final prediction, which we had ignored so far. However, as we were working towards personalized prediction and treatment, we would need to take this information into account, otherwise, we would be at risk of overfitting the data to the trained individuals and would fail to generalize to unseen individuals. For the tumor detection, we had a relatively large patient cohort, however, the number of samples from each patient was small. Interestingly, there was a large overlap in the characteristics of the MRS spectra between the tumor and non-tumor patients. Nevertheless, those non-tumor patients were under the suspicion of having a tumor in the first place.

Sharing some similarities and bearing some differences, these two projects focus on two medical problems that present some commonly present problems when applying ML methods to medical applications. Here, I will summarize the work from each project in detail.

3.2 Early Diagnosis for Epilepsy

In this project, we aimed to gain a better understanding of the mechanism of epileptogenesis and to discover potential EEG biomarkers for epileptogenesis. We conducted experiments using a well-studied rodent model for mesial temporal lobe epilepsy (m-TLE) [68, 69]. From the animal experiments, we collected an enormous amount of longitudinal EEG data from multiple rodents, before and after the epilepsy-inducing electrical stimulation. We did not have detailed annotations except for the recording timestamps. Conducting analysis on such a data set using only human effort is out of the question. DNNs have been demonstrated to have great potential in learning complex features from various data modalities and we wished to exploit this ability to learn from our weakly-supervised EEG data. This was particularly challenging because (1) EEG data is highly non-stationary with relatively low signal-to-noise ratio, (2) it is impossible to obtain detailed annotations on longitudinal EEG recordings, and (3) the underlying epileptogenesis processes are not well-known and complicated with immense cross-subject variability [119].

3.2.1 Summary and Conclusion

In Study I [O1], we explored the possibility of distinguishing EEG signals from before (baseline, BL) and after (epileptogenesis, EPG) the disease-inducing stimulation. As it is suspected that during epileptogenesis the normal brain rhythms could be disrupted [119, 120, 117], which can be reflected in the frequency signature of the EEG recordings, we extracted the frequency components from the EEG traces and fed them into a deep residual neural network for classification. We showed that the network could distinguish between the BL and the EPG

phases accurately, even when trained in a leave-animal-out scheme. This demonstrated a high generalization ability for unseen animals, except for one outlier rat; for this animal, the features learned from other animals indicated the opposite class. We also showed that, indeed, there are certain frequency components that were enhanced or suppressed, which could be indicative of ongoing epileptogenesis. Our findings are consistent with previous work including an increase of the delta rhythm [132, 133, 134, 135, 136], a decrease of the theta rhythm [131, 97], more occurrence of spikes, spike-wave complexes, and sharp waves [76, 75], an increase of high frequency band power [110, 112, 113] in epileptogenesis. Importantly, we showed that the distribution of the DNN prediction was significantly different when pooling over a longer temporal window than just 5 seconds. This indicates a distributional shift of brain activities rather than sudden abrupt changes during epileptogenesis. We showed that one hour of feature aggregation led to very promising discriminative ability of the classifier.

In Study II [O2], we investigated the possibility of learning to distinguish BL and EPG classes with raw EEG data in an end-to-end fashion. We were impressed by the results from DNNs in many tasks, where a great number of details were captured in the network, such as generating high fidelity natural images [5] and high quality human-like voice [38]. We hoped that learning with raw EEG data would allow the network to capture more detailed features for each class from the waveforms directly. In this work, we showed that the DNN can distinguish between the BL and EPG phases by training on five-second EEG segments and, consequently, the problems with the outlier rat in Study I disappeared. This may suggest that the waveforms are better suited for learning with CNN and that the model could further exploit learned features to distinguish between the two classes. Importantly, when the prediction from multiple consecutive segments were aggregated, the performance was seen to improve further. Overall, we obtained an average

AUC of 0.99 across all leave-individual-out cross-validation trials, improved by 0.09 compared to our previous work [O1]. The sensitivity was improved from 0.83 to 0.94 and the specificity was improved from 0.83 to 0.96.

In Study III [O3], we investigated the question of whether the DNN could distinguish the EEG signals from different time windows, i.e., the early stage and the late stage of the epileptogenesis phase. This served as an attempt to stage the progression of epileptogenesis in the latent period, which has never been done before. We demonstrated that the neural network could achieve very high discriminative and generalization abilities in staging epileptogenesis with unseen animals. At the same time, we illustrated that there were features learned by the network that are class-specific and clinically relevant including spike-wave complexes, sharp waves, rhythmic delta rhythm [161, 76, 75]. In addition to the results in terms of classification performance, we also addressed the challenges related to explaining learned features from the network. We explored the network activity and investigated the features that could maximally excite the neurons in the last layer of the network. We demonstrated that units in the last layer respond to different characteristics in the input data, such as spikes, spike-and-waves complexes, spindles, and theta rhythms. Furthermore, there are multiple feature maps are more responsive to different stages of epileptogenesis, which confirms the anticipation that a single biomarker may not be enough to characterize a complex process such as epileptogenesis. A better early diagnosis may rely on a profile of multiple biomarkers [119]. We also explored the class activation map [160] method to visualize the learned features for different classes. We concluded that, with respect to the study for the early diagnosis of epilepsy, we have contributed to advance the current understanding of epileptogenesis in the latent period and have provided evidence that DNNs are capable of learning complex features, without detailed annotations, to detect the presence of epileptogenesis. We have presented some learned EEG features

from the network which resemble features from human patients with hippocampal sclerosis and have paved the way for animal to human translational studies. For instance, we should be able to detect those early-EPG or late-EPG related features in patient recordings and investigate whether they displayed similar trends as in rodent epileptogenesis progression. Furthermore, we could leverage the technique of transfer learning, that transfers the learned features and knowledge, from the rodent epilepsy model to human patients.

3.2.2 Generalizable Insights

ML methods strongly depend on extracting information from a huge amount of data with high quality labels. However, it is often laborious and expensive to obtain expert annotations on massive amounts of data, especially in the medical domain. Thus, this lacking of high quality labels for medical data poses tremendous hurdles for ML application for healthcare. Confronted by this challenge, we utilized a form of "cheap" labelling, i.e., timestamps of the EEG recordings with the relative time to the epilepsy-inducing stimulation in our case. This form of label is cheap and easy to obtain but less informative since during epileptogenesis the epileptogenic abnormalities only happen sparsely and are superimposed on normal brain activities. This leads to a large overlapping of EEG features between both classes. Moreover, to account for the effect of gradual changes of the brain, we propose a prediction aggregation step to pool DNN's output over a long time window. In our work, we show that even in the absence of expert labels of specific EEG waveforms, a large amount of data combined with the "cheap" labels allows us to build a powerful classification system. This way of circumventing the lack of large amount of expert annotations could be applied to a wide range of medical tasks. In other disease staging problems where the progression is gradual and can be better characterized by a distributional shift, our framework could also be

applied.

3.2.3 Limitations

Here, we would like to discuss the limitations in our work and provide some thoughts on several interesting future directions.

In Study I, we could have explored a more specific question when dealing with frequency features, i.e., which frequency band was the most discriminative in discerning the baseline and epileptogenesis phases. Further analysis on why there was an outlier animal, for which the model failed to predict correctly, might also be interesting. When we moved to train the network with the time-series data in Study II, we saw that the issue with the "odd" animal was gone and that the overall performance for unseen animals was significantly higher than that from Study I.

Having being applied to many tasks with sequential data, RNNs may have been a better tool to deal with our EEG recordings so that longer temporal dependencies could be captured by the network. However, as we were learning on individual five-second long EEG segments, the temporal correlation and evolution of EEG patterns captured in these relatively short segments were limited. Although we proposed to use the prediction aggregation method, which seemed to significantly improve the classification performance, this was a *post hoc* analysis. The neural network, itself, remained unaware of the underlying evolution of the signals. We should be able to perform the information aggregation while training, as in [162], where the input could consist of much longer EEG segments and the network would be trained to capture both local features and whole-input level dependencies. The concept of a long-term feature bank, proposed in [163] while dealing with long range video, could also be adapted for capturing long-term patterns and discovering their dependencies.

From the perspective of practical utility, a good biomarker for identifying epileptogenesis in a clinical setting should be noninvasive. In contrast, the data in our study were recorded using a depth electrode, which has a much higher signal-to-noise-ratio compared to surface EEG recordings. In a potential future step, the adaptation of the proposed method to use surface EEG would be of great importance. Furthermore, since in the rodent models the damage to the brain tissue is carefully designed and controlled, it is yet to be verified whether the conclusions still hold for human patients, where the damages are of greater variability. Research on epileptogenesis detection in human patients, compared to that using rodent models, is confronted with even more hurdles such as shorter recording lengths, longer intervals between EEG checkups, and diverse medical conditions.

3.3 Tumor Detection with MRS data

In this project, we tackled the problem of detecting brain tumor samples from the MRS spectra. We were confronted with the challenges that are common in the data-driven healthcare domain, such as noisy labels, data scarcity, and class imbalance.

3.3.1 Summary and Conclusion

In Study IV [O4], we proposed a two-stage framework for tumor detection with noisy labels: (1) a data distillation step to obtain representative samples from both classes and (2) a data augmentation step to generate more samples for training through mixing. We validated the proposed distillation method on the MNIST data set with manually induced label noise and showed that, indeed, the distillation could filter out the samples that were more likely to be cleanly labeled. We also

compared the DNN performance with that of the human neuroradiologists and showed that the network performs on par with the human experts.

In Study V [O5], we extended the previous work by comparing the performance with several other neural network architectures, such as a fully-connected network, an Inception-based [35] network, and a recurrent neural network. We also explored the effects of different strategies during the data augmentation when mixing samples. Furthermore, we explored the method of using a class activation map [160] to inspect the importance weight of each part of the input contributing to the final classification decisions. The results showed that the DNN captured those traditionally considered tumor-typical features, which are consistent with previous studies, including high mI peaks [148], high Cho peak and low Cr peak [147, 149, 150, 155], high Glu-Gln peaks [156, 157], high lactate and lipids peaks [147, 159, 150, 155], and low NAA peak [152, 153, 154, 152]. We showed that our model performs on par with the human experts. However, the limitation of Study IV and V is that each spectrum from each patient was considered as an independent sample during training. The assumption is that samples from each patient have the same label as the patient, which is rarely the case in clinical practice and this induces the label noise. In Study VI [O6], we proposed a framework that uses the multiple instance learning framework (MIL) [56, 55]. This was inspired by the observation that there was a wide distribution of the DNN output for the spectra even from the same patient, which reflected a large disagreement with single spectrum classification. We also observed that not all samples from the same patient show typical class-specific features. However, for tumor patients there should be at least one sample that is tumor-typical. Thus, even with noisy labeling, when more samples from one individual are grouped together, it may be more representative of the true class membership of the patient. To this end, we organized samples in such a way that spectra from each patient were grouped into data bags without the need to label

every single instance in the bag. During training, we applied random sampling and reordering of samples to generate a large number of bags for each patient. However, we still needed to tackle one more problem which was the permutation invariance within the bag, since we did not desire the mapping learned by the network to depend on the order of samples within the bag. To this end, we proposed two modules that could be easily plugged into any network structure to ensure the permutation invariance. We then compared models with and without the permutation invariant modules and showed that our proposed method significantly improved the performance and explainability compared to baselines. Meanwhile, the attention module also provide visualization of how important of each sample in the data bag is to the final classification decidion. From the attention weights, we observed the following: (1) The high attention samples from each class are conventionally considered as representative of that class. For example, in the tumor class, high attention samples are often with high mI peaks [148], high Cho peak and low Cr peak [147, 149, 150, 155], high Glu-Gln peaks [156, 157], high lactate and lipids peaks [147, 159, 150, 155], and low NAA peak [152, 153, 154, 152], all of which have been reported in tumor tissues. (2) Surprisingly, we noticed that some non-tumor patients have spectra with typical tumor-like features, and vice versa, shown in Fig.3 from Study VI. However, with our method, the samples with features from the opposite class are often with the lowest attention weights. (3) From the attention weights, we observed that often samples only contained a subset of the typical metabolic signatures for that class and some features may be more typical than others.

3.3.2 Generalizable Insights

In this project, we tackled multiple issues that are widespread in machine learning for healthcare problems. First, it is common that acquired medical data is anno-

tated at the patient level, e.g., cancer diagnosis [164] or blood cell disorders [54]. However, not all data samples acquired from the patient reflect the pathological abnormalities. Moreover, the size of the data set is often limited by the number of patients and number of samples collected from patients. We proposed a data distillation step and data augmentation step to combat these problems. Furthermore, to get a overview of the patient's data, we proposed to take advantage of the multiple instance learning (MIL) approach [54, 55, 56]. In implementing an MIL framework, we proposed two plug-and-use modules, i.e., an attention and a 3-pooling module. Thus, our approach can be easily transferred to a wide variety of other healthcare applications not limited by the data modality.

3.3.3 Limitations

In this project, there are a few limitations which we would like to address here. In general, we were confronted with two main challenges: (1) a lack of training data and (2) noisy labeling. The first problem was also reflected by a large variance during training which is shown by large performance differences in the training and the validation sets. In this case, the limited size of the training set was a critical contributing factor. Hence, it is possible that effective data augmentation approaches could improve the performance. We implemented one of the most straightforward methods, i.e., mixing samples, in order to increase the number of training samples. However, the simple mixing did not improve the performance by a large margin.

In order to tackle the second problem, we proposed the "distillation" step, which enabled the selection of some stereotypical samples for each class. However, there exists a trade-off between how typical the selected samples can be and the number of selected samples. On the one hand, a strict selection in the distillation will likely yield a small number of samples, which will not be sufficient to represent

diverse patterns. On the other hand, when the distillation is loose, we would end up with a big set of data that are similarly noisy. Hence, more explorations to find a way to balance the distillation step would be of interest. In Study IV and V, we considered each spectrum from the same patient individually, omitting the fact that they were from the same entity. In Study V, we explored this idea and proposed the classification based on MIL [56, 55]. However, the generation of the training data is limited by the simple oversampling and random reordering method. For patients who have an extremely small number of samples, the generated data bags will be filled with too many copies of the same spectrum, which is not ideal. This could be addressed with a stratified oversampling strategy, i.e., the more single spectra one patient has, the more bags we generate for this patient. On the other hand, the model will be biased towards patients with many bags of data. Further experiments implementing other data augmentation strategies are, therefore, very much desirable.

3.4 Outlook

3.4.1 Early diagnosis of epilepsy

There are numerous interesting directions that we could pursue further for the ultimate purpose of improving individualised diagnoses and personalized disease progression trajectory predictions. To this end, we would need to acquire data from different animal models as well as from human patients. It would be of great significance to discover robust and reliable biomarkers across different animal epilepsy models and then to translate these findings to human patients.

Clinically, we are interested in knowing for how long it is necessary to record the EEG from patients in order to yield a reliable prediction and, in addition, how

often the patients need to be EEG monitored in order to track the key points of the progression of epilepsy and so not to miss the optimal intervention window.

By approaching the early diagnosis of epilepsy from the supervised learning direction, we are limited, to some extent, by the quality of the labels. However, addressing this problem in an unsupervised fashion could provide us with more insights. For example, brain diseases aside, we could train an encoder on an enormous amount of data from large groups of healthy and non-healthy individuals such that the model could learn robust and comprehensive representations of EEG signals in a diverse set of conditions for animals as well as humans. In this way, we could build an activation atlas for EEG signals similar to the atlas for images in [165]. We could also extract information that is individual-specific, such as an identity vector in speaker identification in the speech signal processing domain. This information could then be used to customize the prediction for unseen subjects in the future. We would expect that there may be a distributional shift of the EEG representation during the progress of epilepsy for each individual and for different brain disorders, as they progress differently in the representation space. In this way, we could not only obtain the disease progression trajectory for epilepsy, but also other neurological disorders.

Moreover, it would be interesting to combine the findings from the machine learning perspective with studies the computational brain circuitry modeling. Concerning EEG signals, we can only record the population activity without anatomical knowledge, thus, it is difficult to infer the structure and function at the single-neuron level. Computational modeling provides a fine-grained monitor and understanding of what is happening with the underlying circuit. Combining both lines of research could potentially provide us with a better understanding of the mechanism of epileptogenesis.

3.4.2 Brain Tumor Detection with MRS data

For future work on brain tumor detection with MRS data, we could address the noisy labeling from the active learning point of view, which focuses on selecting "good" samples for training. Such a line of thinking could be implemented based on the sample's contribution to the training gradient [166], or its contribution to the training loss [167, 168], or through the area under the margin ranking to identify the mislabeled data [169]. Effective data augmentation methods are also beneficial to expand the training data, as well as to cover variability introduced by patients individual conditions, such as Mixup [170] and creating samples from generative models [171].

In this thesis, I have shown how ML methods can be successfully applied to two medical problems. We hope our work will lead to improved clinical practice and ultimately help patients and inspire new research directions.

Kapitel 4

Deutsche Zusammenfassung

Die Methoden des maschinellen Lernens (ML) waren in den vergangenen Jahren sehr erfolgreich und haben ihr großes Potential in vielen Forschungsgebieten gezeigt, z. B. das Lernen von Spielen [2, 3, 4], das Generieren hochwertiger Bilder [5, 6], Style Transfer [7], Spracherkennung und Synthese [8, 9] sowie die Verarbeitung natürlicher Sprache [10, 11]. Das maschinelle Lernen profitiert stark von der immer größeren Rechenleistung, der Verfügbarkeit großer und spezialisierter Datensätze und tieferen theoretischen Einsichten in viele Lernalgorithmen.

In den letzten Jahren gab es eine Vielzahl von Forschungbemühungen, die sich mit der Anwendung von ML-Methoden im Gesundsheitsbereich befassen. Es gibt beeindruckende Arbeiten für diverse medizinische Probleme, beispielsweise Klassifikation von Herzkreislauferkrankungen [12], Hautkrebserkennung [13], Lungenkrebsdiagnose [14], automatische Vorhersage von Erkrankungen [15], sowie COVID-19 Diagnose und Behandlung [16].

In dieser Dissertation befassen wir uns mit dem Problem, ML-methoden im Kontext von Epilepsie- und Gehirntumorerkennung anzuwenden. Im ersten Projekt versuchen wir den Krankheitsverlauf in der latenten Epileptogenesephase (nach der Gehirnschädigung aber vor dem ersten spontanen epileptischen Anfall) zu ver-

stehen und frühzeitig Vorhersagen zu treffen, ob ein bestimmtes Individuum ein hohes Risiko hat, Epilepsie zu entwickeln oder nicht. Im zweiten Projekt zielen wir darauf ab, ein Pre-Screening-Werkzeug zu entwickeln, welches Gehirntumore basierend auf Magnetresonanzspektroskopiedaten (MRS–Daten) erkennen kann. Im Folgenden werden wir unsere Arbeit zu diesen beiden Themen zusammenfassen.

4.1 Frühdiagnose von Epilepsie

Für dieses Projekt haben unsere Kooperationspartner vom Universitätsklikium Frankfurt Experimente mit einem gut erforschten Nagetiermodell für Epilepsie des mesialen Temporallappens (mesial temporal lobe epilepsie, m-TLE) durchgeführt [68]. Von den Tierversuchen haben wir eine große Menge longitudonaler EEG–Daten von mehreren Nagetieren vor und nach der epilepsieauslösenden elektrischen Stimulation gesammelt. Wir verfügen dabei nicht über detaillierte Annotationen außer den Aufnahmezeitstempeln, welche eine Form von schwachen Labels darstellen. Tiefe neuronale Netze (deep neural networks, DNNs) haben ein hervorragendes Potential gezeigt, komplexe Features aus verschiedenen Datenmodalitäten zu lernen, und wir möchten diese Fähigkeit nutzen, um aus unseren schwach annotierten EEG–Daten zu lernen. Dies ist besonders herausfordernd, denn 1. sind EEG–Signale in hohem Maße nicht stationär mit einem relativ niedrigen Signal-Rausch-Verhältnis, 2. ist der Epileptogeneseprozess nicht gut verstanden und es ist nicht möglich, detaillierte Annotationen der logitudonalen EEG Aufnahmen zu erstellen, und 3. gibt es eine enorme Variablität zwischen den Individuen in den EEG–Aufnahmen.

In Paper I versuchen wir, EEG–Signale aus zwei Phasen zu unterscheiden: vor (baseline, BL) und nach (Epileptogenese, EPG) der Stimulation, durch welche die Erkrankung ausgelöst wird. Es wird vermutet, dass während der Epileptogene-

se der normale Gehirnrhythmus gestört wird, was sich im Frequenzspektrum der EEG–Aufnahmen wiederspiegelt. Wir haben deshalb die Frequenzanteile aus den EEG–Signalen extrahiert und als Eingabe für ein tiefes neuronales Netzwerk mit residualen Verbindungen genutzt. So haben wir gezeigt, dass das Netzwerk die BL– und EPG–Phasen zuverlässig unterscheiden konnte, sogar wenn ein Trainingsschema zum Einsatz kam, bei dem die Daten eines Individuums während des Trainings komplett vorenthalten wurden. Dies zeigt eine gute Generalisierungsfähigkeit auf zuvor nicht gesehene Tiere, abgesehen von einer einzelnen unzypischen Ratte. Wir haben außerdem gezeigt, dass es in der Tat bestimmte Frequenzanteile gibt, die während der Epileptogenese verstärkt oder unterdrückt werden.

Statt das Frequenzspektrum zu nutzen, untersuchten wir im Paper II die Möglichkeit, mit rohen EEG Daten Ende-zu-Ende zu trainieren, um die BL– und EPG–Phasen zu unterscheiden. Wir waren von den Ergebnissen von DNNs bei vielen Aufgaben beeindruckt, bei denen viele Details vom Netzwerk erfasst werden können, z. B. beim Erzeugen von realistischen natürlichen Bildern [5] und beim Generieren von qualitativ hochwertiger menschenähnlicher Stimmen [38]. Daher hofften wir, dass das Netz beim Lernen auf rohen EEG–Daten detailliertere Features direkt in den Wellenformen erfassen könnte. In dieser Arbeit zeigten wir, dass ein DNN die BL– und EPG–Phasen basierend auf fünf Sekunden langen EEG Segmenten unterscheiden kann, und dass die Probleme mit der untypischen Ratte aus Paper I nicht mehr auftreten. Diese Ergebnisse könnten nahelegen, dass Wellenformen besser für das Lernen mit Convolutional Neural Networks (CNNs) geeignet sind und das Modell weitere gelernte Features nutzen konnte, um die beiden Klassen zu unterscheiden. Die Klassifikationsergebnisse können zudem weiter verbessert werden, indem die Prädiktionen von mehreren aufeinander folgenden Segmenten aggregiert werden.

In Paper III untersuchten wir schließlich die Frage, ob ein DNN EEG–Signale

aus verschiedenen Zeitfenstern unterscheiden kann; konkret betrachteten wir dabei die frühe Phase und die späte Phase der Epileptogenese. Dies kann als Versuch gesehen werden, das Fortschreiten der Epileptogenese in der latenten Periode zu quantifizieren. Wir haben gezeigt, dass das so trainierte neuronale Netz über sehr gute diskriminative Fähigkeiten verfügt und auch auf zuvor nicht gesehene Tiere generalisieren kann. Außerdem befassten wir uns mit der Frage, wie man die gelernten Features erklären kann. Dazu untersuchten wir die Netzwerkaktivität und analysierten die Features, welche zu einer maximalen Aktivität der Neuronen in der letzten Schicht des Netzes führen. Wir zeigten, dass die Neuronen in der letzten Schicht auf bestimmte Charakteristiken in den Eingabedaten reagieren, z. B. Spikes, Spike-and-waves, Spindeln sowie Thetarhythmen. Die Class–Activation–Mapping–Methode wurde außerdem eingesetzt, um gelernte Features für die verschiedenen Klassen zu visualisieren.

Zusammengefasst lässt sich sagen, dass wir zu einem tieferen Verständnis der Epileptogenese in der latenten Phase beigetragen haben und gezeigt haben, dass tiefe neuronale Netze, mit denen das Auftreten der Epileptogenese erkannt werden kann, in der Lage sind, komplexe Features auch ohne detaillierte Annotationen zu lernen. Wir haben einige vom Netz gelernte Features präsentiert, welche Features der Hippocampussklerose beim Menschen ähneln, und damit den Weg für Studien zur Übertragung vom Tier auf den Menschen geebnet. Zum Beispiel könnte man die Features der frühen oder späten EPG–Phase in Patientenaufnahmen erkennen und untersuchen, ob sich ähnliche Trends wie in der Entwicklung der Epileptogenese bei Nagetieren zeigen. Des Weiteren könnte man Methoden des Tranfer Lernings nutzen, um gelerntes Wissens aus dem Nagetierepilepsiemodell auf menschliche Patienten zu übertragen.

Einschränkungen und Erweiterungsmöglichkeiten

In Paper I könnten wir weiter untersuchen, welche Frequenzbänder die beste Unterscheidung von der BL– und EPG–Phasen erlauben. Außerdem wäre eine genauere Untersuchung unteressant, wieso es ein untypisches Tier gibt, bei dem das Model nicht in der Lage ist, die Phase korrekt zu prädizieren. Als wir in Paper II dazu übergingen, das Netz basierend auf Zeitreihendaten zu trainieren, sahen wir, dass Probleme mit diesem untypischen Tier verschwanden und auch die generelle Genauigkeit bei zuvor nicht gesehenen Tieren deutlich höher lag als in Paper I.

Da rekurrente neuronale Netze (RNNs) häufig für Aufgaben mit sequentiellen Daten eingesetzt werden, könnten sie ein besseres Werkzeug sein, um EEG–Aufnahmen zu verarbeiten, so dass längere zeitliche Abhängigkeiten vom Netz erfasst werden könnten. Da wir bisher mit einzelnen EEG–Segmenten mit einer Länge von fünf Sekunden trainieren, kann nur eine stark begrenzte zeitliche Korrelation betrachtet werden. Auch wenn wir eine Aggregationsmethode für Prädiktionen vorgeschlagen haben, welche die Klassifikationsgenauigkeit deutlich verbessert, handelt es sich dabei nur um eine *post-hoc* Analyse. Das neuronale Netzwerk selbst ist sich der zugrunde liegenden Entwicklung der Signale nicht bewusst. Wir könnten die Informationsaggregation bereits während des Trainings durchführen wie in [162], wo die Eingabe aus viel längeren Segmenten besteht und das Netzwerk sowohl lokale Features als auch Zusammenhänge über die gesamte Eingabe hinweg erfassen kann. Das Konzept einer langfristigen Feature–Bank, das in [163] vorgeschlagen wurde, um lange Videosequenzen zu verarbeiten, könnte ebenfalls adaptiert werden, um langfristige Muster zu erfassen und deren Abhängigkeiten zu entdecken.

Als nächsten Schritt wäre es wichtig, die vorgestellten Methoden auf Oberflächen–EEG zu übertragen. Da in den Nagetiermodellen die Schädigung des Gehirns sorgsam geplant und kontrolliert wurde, muss zudem noch überprüft werden, ob die

Schlussfolderungen auch für menschliche Patienten gelten, wo die Schädigungen stärker variieren. Bei der Erfoschung der EPG bei menschlichen Patienten gibt es mehr Hürden als bei Nagetiermodellen, z. B. eine kürzere Aufnahmedauer, längere Intervalle zwischen EEG–Untersuchungen, und vielfältige Krankheitsbilder.

4.2 Tumorerkennung mit MRS–Daten

In diesem Abschnitt werden wir unsere Arbeiten zum Thema der Tumorerkennung basierend auf MRS–Daten zusammenfassen. In diesen Projekt befassen wir uns mit typischen Herausforderungen datengetriebener medizinischer Anwendungen so wie verrauschten Labels, Datenknappheit und Klassenungleichgewicht.

In Paper IV [O4] schlugen wir ein Framwork zur Tumordetektion mit verrauschten Labels vor, welches aus zwei Schritten besteht, 1) ein Datendestillationsschritt, um repräsentative Samples von beiden Klassen zu erhalten, und 2) ein Datenaugmentierungsschritt, um durch Mischen zusätzliche Samples zu generieren. Wir haben die vorgeschlagene Destillationsmethode mit dem MNIST–Datensatz mit manuell induziertem Rauschen in den Labels validiert und gezeigt, dass die Destillation tatsächlich Samples herausfiltern konnte, die eine höhere Wahrscheinlichkeit haben, korrekte gelabelt zu sein. Die vorgestellte Methode erzielte vergleichbar gute Ergebnisse wie menschliche Experten.

In Paper V [O5] erweiterten wir unsere vorherige Arbeit, indem wir mehrere Achitekturen für das neuronale Netz hinsichtlich ihrer Klassifikationsgenauigkeit verglichen. Dabei betrachteten wir ein vollständig verbundenes neuronales Netz, ein Inception-basiertes Netz und ein rekurrentes neuronales Netz. Außerdem erforschten wir den Einfluss verschiedener Datenaugmentierungsstrategien beim Mischen von Samples. Zudem untersuchten wir mit der Class–Activation–Map–

Methode [160], wie stark unterschiedliche Teile der Eingabe zum Klassifikationsergebnis beitragen.

Um einen größeren Überblick über die Daten jedes Patienten zu erlangen, schlugen wir anschließend in Paper VI [O6] ein Framework vor, das auf Multiple Instance Learning (MIL) basiert. Dieser Ansatz ist von der Beobachtung inspiriert, dass nicht alle Samples eines Patienten zwangsläufig klassentypische Features zeigen. Allerdings sollte es bei Patienten mit Tumoren mindestens ein Sample geben, welches typisch für einen Tumor ist. Daher könnten mehrere gruppierte Samples eines Patienten trotz der verrauschten Labels repräsentativer für die tatäschliche Klasse des Patienten sein. Darum gruppierten wir Samples von jedem Patienten in sogenante Bags, wofür wir keine Labels für jedes einzelne Sample innerhalb der Bag mehr benötigen. Wir schlugen zwei Module vor, die leicht in beliebige Netzstrukturen integriert werden können, um Permutationsinvarianz innerhalb der Bags zu erzielen. Anschließend führten wir eine Ablationsstudie unserer vorgeschlagenen Methode durch und zeigten, dass das vorgestellte Verfahren signifikant bessere Ergebnisse als voherige Verfahren erzielt und zudem besser erklärbar ist.

Einschränkungen und Erweiterungsmöglichkeiten

In diesem Projekt gibt es einige Einschränkungen, die wir hier diskutieren möchten. Generell stehen wir primär zwei Herausforderungen gegenüber, 1) dem Mangel an Trainingsdaten, und 2) verrauschten Labels. Das erste Problem führt zu einer größen Varianz während des Trainings, was sich darin äußert, dass es einen großen Unterschied zwischen der Klassifikationsgenauigkeit auf dem Trainings- und dem Validierungsdatensatz gibt. In diesem Fall ist die beschränkte größe des Trainingsdatensatzes ein kritischer Faktor. Deshalb könnten effektive Datenaugmentierungsverfahren die Ergebnisse weiter verbessern. Wir haben eine der einfachsten Methoden, das Mischen von Samples, implementiert, um die Anzahl der Trainingsbeispie-

le zu erhöhen. Allerdings hat das einfache Mischen nicht zu einer großen Verbesserung geführt. Um das zweite Problem zu lösen, haben wir einen Destillationsschritt vorgeschlagen, der typische Samples für jede Klasse auswählt. Allerdings gibt es einen Trade-Off zwischen der Anzahl der gewählten Samples und darin, wie typisch sie für die betrachtete Klasse sind. Einerseits liefert eine strenge Auswahl in der Destillation nur eine geringe Menge an Daten, die nicht ausreicht, um vielfältige Muster zu repräsentieren. Andererseits führt eine zu zu großzügige Destillation zu einen großen Datensatz, der genauso verrauscht ist wie die ursprünglichen Daten. Deshalb ist es von Interesse, genauer zu untersuchen, wie man für die Destillation eine gute Balance zwischen beiden Extremem finden kann. In Paper IV und Paper V haven die Spektra eines Patienten einzeln betrachtet, ohne zu berücksichtigen, dass sie vom gleichen Patienten stammen. In Paper V haven wir dies genauer Untersucht und Klassifikation mittels MIL vorgeschlagen. Allerdings werden die Trainingsdaten nur mit einer einfachen Oversampling-Methode generiert. Für Patienten mit sehr wenigen Samples enthalten die Bags viele Kopien des selben Spektrums, was nicht ideal ist. Mit einer stratifizierten Oversampling-Strategie, bei der umso mehr Bags erzeugt werden je mehr Spektra von dem jeweiligen Patienten vorliegen, würde das Modell allerdings einen Bias für Patienten mit vielen Daten erhalten. Weitere Experimente mit anderen Augmentierungsstrategien sind darum wünschenswert.

4.3 Ausblick

4.3.1 Frühdiagnose von Epilepsie

Es gibt viele interessante Forschungsrichtungen, die wir in der Zukunft verfolgen könnten, um den langfristigen Ziel näherzukommen, individaulisierte Diagonosen zu erstellen und die persönliche Entwicklung der Erkankung vorherzusagen. Um

dieses Ziel zu erreichen, müssten wir Daten von verschiedenen Tiermodellen und menschlichen Patienten erfassen. Es wäre von großer Bedeutung, robuste und zuverlässige Biomarker zu finden, die sich über mehrere verschiedene Tieremodelle für Epilepsie verallgemeinern lasssen, und diese dann auf den Menschen zu übertragen.

Aus klinischer Sicht interessiert uns die Frage, wie lange einzelne EEG–Aufnahmen sein müssen, um eine zuverlässige Vorhersage zu erlauben, und wie häufig EEGs der Patienten aufgezeichnet werden müssen, um die Schlüsselpunkte der Entwicklung der Epilepsie zu verfolgen und das optimale Interventionsfenster nicht zu verpassen. Wenn man die Frühdiagnose aus der Perspektive des überwachten Lernens betrachtet, ergeben sich Einschränkungen durch die Qualität der Labels. Daher könnte es neue Einsichten liefern, das Problem aus der Sicht des unüberwachten Lernens zu betrachten. Lässt man Erkrankungen des Gehirns einmal außen vor, könnte man beispielsweise einen Encoder basierend auf riesigen Datenmengen von großen Gruppen gesunder und kranker Individuen trainieren, so dass das Modell eine robuste und umfassende Repräsentation von EEG–Signalen unter vielfältigen Bedingungen für Tiere wie auch für Menschen lernnen könnte. Auf diesem Weg könnten wir einen Aktivierungsatlas wie in [165] aufbauen, aber nicht für Bilder sondern für EEG–Signale. Zudem könnten wir Informationen extrahieren, die spezifisch für ein bestimmtes Individuum sind. Dies ist ähnlich zum Identitätsvektor aus dem Bereich der Sprecheridentifikation in der Sprachsignalverarbeitung. Auf Grundlage dieser Information könnten wir dann die Prädiktion für zuvor nicht gesehene Individuen anpassen. Man würde annehmen, dass es in jedem Individuum verschiedene Verteilungsverschiebungen der EEG–Repräsentation während der Entwicklung der Epilepsie gibt und dass sich verschiedene Störungen im Gehirn im Raum der Repräsentationen unterschiedlich entwickeln. Auf diesem Weg könnte man nicht nur die Krankheitsentwicklung für Epilepsie erfassen sondern auch

die anderer neurologischer Störungen.

Außerdem wäre es interessant, die Erkenntnisse aus der Perspektive des maschinellen Lernens mit Arbeiten zu kombinieren, welche die Berechnungen der neuronalen Schaltkreise modellieren. Mit EEG–Signalen können wir nur die Aktivität einer ganzen Population aufnehmen und es ist schwierig Aussagen auf der Stufe eines einzelnen Neurons zu treffen. Berechnungsmodelle bieten ein feingranulareres Verständnis, was in den zugrunde liegenden Schaltkreisen vor sich geht. Eine Kombination dieser beiden Forschungsrichtungen könnte uns potenziell ein besseres Verständnis der Mechanismen der Epileptogenese erlauben.

4.3.2 Gehirntumordetektion mit MRS-Daten

In zukünftigen Arbeiten zur Gehirntumordetektion mit MRS–Daten könnten wir die verrauschten Labels aus der Perspektive des aktiven Lernens betrachten. Dies ist darauf fokussiert, gute Samples für das Training auszuwählen. Man könnte so ein Verfahren implementieren, indem man den Beitrag jedes Samples zum Gradienten während des Trainings [166] oder der Loss-Funktion [167, 168] betrachtet. Alternativ ließe sich auch das sogenannte *area under the margin ranking* verwenden, um falsch gelabelte Daten zu identifizieren [169]. Effektive Datenaugmentierungsmethoden wären ebenfalls vorteilhaft, um die Trainingsdaten zu erweitern und die Variabilität der verschiedenen Patienten mit unterschiedlichen Krankheitsbildern abzudecken. Konkret wäre es denkbar, Mixup [170] zu nutzen oder Samples mit Hilfe von generativen Modellen zu erzeugen [171].

Bibliography

[1] A. Pitkänen, "Therapeutic approaches to epileptogenesis—hope on the horizon," *Epilepsia*, vol. 51, pp. 2–17, 2010.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[3] I. Szita and A. Lörincz, "Learning tetris using the noisy cross-entropy method," *Neural computation*, vol. 18, no. 12, pp. 2936–2941, 2006.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[5] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[6] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

[7] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

[8] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964, IEEE, 2016.

[9] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, pp. 1764–1772, PMLR, 2014.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[14] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[15] A. Alaa and M. Schaar, "Autoprognosis: Automated clinical prognostic modeling via bayesian optimization with structured kernel learning," in *International conference on machine learning*, pp. 139–148, PMLR, 2018.

[16] M. Jamshidi, A. Lalbakhsh, J. Talla, Z. Peroutka, F. Hadjilooei, P. Lalbakhsh, M. Jamshidi, L. La Spada, M. Mirmozafari, M. Dehghani, *et al.*, "Artificial intelligence and covid-19: deep learning approaches for diagnosis and treatment," *IEEE Access*, vol. 8, pp. 109581–109595, 2020.

[17] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, p. e7, 2017.

[18] D. A. Drachman, "Do we have brain to spare?," 2005.

[19] G. G. Turrigiano and S. B. Nelson, "Homeostatic plasticity in the developing nervous system," *Nature reviews neuroscience*, vol. 5, no. 2, pp. 97–107, 2004.

[20] M. P. Walker, "The role of slow wave sleep in memory processing," *Journal of Clinical Sleep Medicine*, vol. 5, no. 2 suppl, pp. S20–S26, 2009.

[21] V. Ego-Stengel and M. A. Wilson, "Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat," *Hippocampus*, vol. 20, no. 1, pp. 1–10, 2010.

[22] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[24] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," *Nature Reviews Neuroscience*, vol. 21, no. 6, pp. 335–346, 2020.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[28] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[29] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[37] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *arXiv preprint arXiv:1802.06955*, 2018.

[38] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[39] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[40] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling," in *Thirteenth annual conference of the international speech communication association*, 2012.

[41] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen, "Breaking the softmax bottleneck: A high-rank rnn language model," *arXiv preprint arXiv:1711.03953*, 2017.

[42] S. Takase, J. Suzuki, and M. Nagata, "Character n-gram embeddings to improve rnn language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5074–5082, 2019.

[43] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[44] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10971–10980, 2020.

[45] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649, Ieee, 2013.

[46] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *arXiv preprint arXiv:1706.02737*, 2017.

[47] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[49] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[50] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, pp. 1310–1318, PMLR, 2013.

[51] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[52] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[53] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[54] A. Sadafi, A. Makhro, A. Bogdanova, N. Navab, T. Peng, S. Albarqouni, and C. Marr, "Attention based multiple instance learning for classification of blood cell disorders," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 246–256, Springer, 2020.

[55] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *CoRR*, vol. abs/1802.04712, 2018.

[56] G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao, "Multiple instance learning for computer aided diagnosis," *Advances in neural information processing systems*, vol. 19, p. 425, 2007.

[57] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *Neural Networks*, vol. 117, pp. 8–66, 2019.

[58] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *2017 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0588–0592, IEEE, 2017.

[59] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[60] A. Ng, "Machine learning yearning," *URL: http://www. mlyearning. org/(96)*, 2017.

[61] A. Pitkänen and J. Engel, "Past and present definitions of epileptogenesis and its biomarkers," *Neurotherapeutics*, vol. 11, no. 2, pp. 231–241, 2014.

[62] F. L. Da Silva, "Eeg: origin and measurement," in *EEg-fMRI*, pp. 19–38, Springer, 2009.

[63] S. P. Jadhav, C. Kemere, P. W. German, and L. M. Frank, "Awake hippocampal sharp-wave ripples support spatial memory," *Science*, vol. 336, no. 6087, pp. 1454–1458, 2012.

[64] M. F. Carr, S. P. Jadhav, and L. M. Frank, "Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval," *Nature neuroscience*, vol. 14, no. 2, p. 147, 2011.

[65] M. R. Mehta, "Cortico-hippocampal interaction during up-down states and memory consolidation," *Nature neuroscience*, vol. 10, no. 1, pp. 13–15, 2007.

[66] M. György Buzsáki, *The brain from inside out*. Oxford University Press, 2019.

[67] K. S. Anand and V. Dhikav, "Hippocampus in health and disease: An overview," *Annals of Indian Academy of Neurology*, vol. 15, no. 4, p. 239, 2012.

[68] B. A. Norwood, S. Bauer, S. Wegner, H. M. Hamer, W. H. Oertel, R. S. Sloviter, and F. Rosenow, "Electrical stimulation-induced seizures in rats: a "dose-response" study on resultant neurodegeneration," *Epilepsia*, vol. 52, no. 9, pp. e109–e112, 2011.

[69] L. S. Costard, V. Neubert, M. T. Venø, J. Su, J. Kjems, N. M. Connolly, J. H. Prehn, G. Schratt, D. C. Henshall, F. Rosenow, *et al.*, "Electrical stimulation of the ventral hippocampal commissure delays experimental epilepsy and is associated with altered microrna expression," *Brain Stimulation*, vol. 12, no. 6, pp. 1390–1401, 2019.

[70] I. Vida, "Morphology of hippocampal neurons," in *Hippocampal Microcircuits*, pp. 27–67, Springer, 2010.

[71] L. L. Colgin, "Rhythms of the hippocampal network," *Nature Reviews Neuroscience*, vol. 17, no. 4, pp. 239–249, 2016.

[72] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *science*, vol. 304, no. 5679, pp. 1926–1929, 2004.

[73] G. Buzsáki and B. O. Watson, "Brain rhythms and neural syntax: implications for efficient coding of cognitive content and neuropsychiatric disease.," *Dialogues in clinical neuroscience*, vol. 14, no. 4, p. 345, 2012.

[74] M. Vandecasteele, V. Varga, A. Berényi, E. Papp, P. Barthó, L. Venance, T. F. Freund, and G. Buzsáki, "Optogenetic activation of septal cholinergic neurons suppresses sharp wave ripples and enhances theta oscillations in the hippocampus," *Proceedings of the National Academy of Sciences*, vol. 111, no. 37, pp. 13535–13540, 2014.

[75] G. Buzsáki, "Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning," *Hippocampus*, vol. 25, no. 10, pp. 1073–1188, 2015.

[76] M. Valero, R. G. Averkin, I. Fernandez-Lamo, J. Aguilar, D. Lopez-Pigozzi, J. R. Brotons-Mas, E. Cid, G. Tamas, and L. M. de la Prida, "Mechanisms for selective single-cell reactivation during offline sharp-wave ripples and their distortion by fast ripples," *Neuron*, vol. 94, no. 6, pp. 1234–1247, 2017.

[77] A. C. Singer, M. F. Carr, M. P. Karlsson, and L. M. Frank, "Hippocampal swr activity predicts correct decisions during the initial learning of an alternation task," *Neuron*, vol. 77, no. 6, pp. 1163–1173, 2013.

[78] G. Girardeau, K. Benchenane, S. I. Wiener, G. Buzsáki, and M. B. Zugaro, "Selective suppression of hippocampal ripples impairs spatial memory," *Nature neuroscience*, vol. 12, no. 10, pp. 1222–1223, 2009.

[79] Y. Liu, S. S. McAfee, and D. H. Heck, "Hippocampal sharp-wave ripples in awake mice are entrained by respiration," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[80] G. T. Neske, "The slow oscillation in cortical and thalamic networks: mechanisms and functions," *Frontiers in neural circuits*, vol. 9, p. 88, 2016.

[81] N. W. Schultheiss, M. Schlecht, M. Jayachandran, D. R. Brooks, J. L. McGlothan, T. R. Guilarte, and T. A. Allen, "Awake delta and theta-rhythmic hippocampal network modes during intermittent locomotor behaviors in the rat.," *Behavioral Neuroscience*, 2020.

[82] G. Buzsáki, "Theta oscillations in the hippocampus," *Neuron*, vol. 33, no. 3, pp. 325–340, 2002.

[83] Y. Isomura, A. Sirota, S. Özen, S. Montgomery, K. Mizuseki, D. A. Henze, and G. Buzsáki, "Integration and segregation of activity in entorhinal-hippocampal subregions by neocortical slow oscillations," *Neuron*, vol. 52, no. 5, pp. 871–882, 2006.

[84] V. V. Vyazovskiy, U. Olcese, E. C. Hanlon, Y. Nir, C. Cirelli, and G. Tononi, "Local sleep in awake rats," *Nature*, vol. 472, no. 7344, pp. 443–447, 2011.

[85] A.-M. Costa, C. Lucchi, A. Malkoç, C. Rustichelli, and G. Biagini, "Relationship between delta rhythm, seizure occurrence and allopregnanolone hippocampal levels in epileptic rats exposed to the rebound effect," *Pharmaceuticals*, vol. 14, no. 2, p. 127, 2021.

[86] G. Pellegrino, A. Machado, N. von Ellenrieder, S. Watanabe, J. A. Hall, J.-M. Lina, E. Kobayashi, and C. Grova, "Hemodynamic response to interictal epileptiform discharges addressed by personalized eeg-fnirs recordings," *Frontiers in Neuroscience*, vol. 10, p. 102, 2016.

[87] H. Nariai, N. Matsuzaki, C. Juhász, T. Nagasawa, S. Sood, H. T. Chugani, and E. Asano, "Ictal high-frequency oscillations at 80–200 hz coupled with delta phase in epileptic spasms," *Epilepsia*, vol. 52, no. 10, pp. e130–e134, 2011.

[88] U. Sławińska and S. Kasicki, "The frequency of rat's hippocampal theta rhythm is related to the speed of locomotion," *Brain research*, vol. 796, no. 1-2, pp. 327–331, 1998.

[89] G. Buzsáki and E. I. Moser, "Memory, navigation and theta rhythm in the hippocampal-entorhinal system," *Nature neuroscience*, vol. 16, no. 2, pp. 130–138, 2013.

[90] S. Amemiya and A. D. Redish, "Hippocampal theta-gamma coupling reflects state-dependent information processing in decision making," *Cell reports*, vol. 22, no. 12, pp. 3328–3338, 2018.

[91] G. R. Richard, A. Titiz, A. Tyler, G. L. Holmes, R. C. Scott, and P.-P. Lenck-Santini, "Speed modulation of hippocampal theta frequency correlates with spatial memory performance," *Hippocampus*, vol. 23, no. 12, pp. 1269–1279, 2013.

[92] J. O'Keefe and M. L. Recce, "Phase relationship between hippocampal place units and the eeg theta rhythm," *Hippocampus*, vol. 3, no. 3, pp. 317–330, 1993.

[93] W. E. Skaggs, B. L. McNaughton, M. A. Wilson, and C. A. Barnes, "Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences," *Hippocampus*, vol. 6, no. 2, pp. 149–172, 1996.

[94] S. M. Montgomery, A. Sirota, and G. Buzsáki, "Theta and gamma coordination of hippocampal networks during waking and rapid eye movement sleep," *Journal of Neuroscience*, vol. 28, no. 26, pp. 6731–6741, 2008.

[95] C. Müller and S. Remy, "Septo–hippocampal interaction," *Cell and tissue research*, vol. 373, no. 3, pp. 565–575, 2018.

[96] E. Kropff, J. E. Carmichael, E. I. Moser, and M.-B. Moser, "Frequency of theta rhythm is controlled by acceleration, but not speed, in running rats," *Neuron*, vol. 109, no. 6, pp. 1029–1039, 2021.

[97] L. Chauviere, N. Rafrafi, C. Thinus-Blanc, F. Bartolomei, M. Esclapez, and C. Bernard, "Early deficits in spatial memory and theta rhythm in experimental temporal lobe epilepsy," *Journal of Neuroscience*, vol. 29, no. 17, pp. 5402–5410, 2009.

[98] A. K. Engel, D. Senkowski, and T. R. Schneider, "Multisensory integration through neural coherence," *The neural bases of multisensory processes*, 2012.

[99] A. L. Lockmann, D. A. Laplagne, and A. B. Tort, "Olfactory bulb drives respiration-coupled beta oscillations in the rat hippocampus," *European Journal of Neuroscience*, vol. 48, no. 8, pp. 2663–2673, 2018.

[100] A. Bibbig, S. Middleton, C. Racca, M. J. Gillies, H. Garner, F. E. LeBeau, C. H. Davies, and M. A. Whittington, "Beta rhythms (15–20 hz) generated by nonreciprocal communication in hippocampus," *Journal of neurophysiology*, vol. 97, no. 4, pp. 2812–2823, 2007.

[101] L. M. Rangel, A. A. Chiba, and L. K. Quinn, "Theta and beta oscillatory dynamics in the dentate gyrus reveal a shift in network processing state during cue encounters," *Frontiers in systems neuroscience*, vol. 9, p. 96, 2015.

[102] S. Iwasaki, T. Sasaki, and Y. Ikegaya, "Hippocampal beta oscillations predict mouse object-location associative memory performance," *Hippocampus*, vol. 31, no. 5, pp. 503–511, 2021.

[103] A. S. França, G. C. do Nascimento, V. Lopes-dos Santos, L. Muratori, S. Ribeiro, B. Lobão-Soares, and A. B. Tort, "Beta2 oscillations (23–30 hz) in the mouse hippocampus during novel object recognition," *European Journal of Neuroscience*, vol. 40, no. 11, pp. 3693–3703, 2014.

[104] M. Bartos, I. Vida, and P. Jonas, "Synaptic mechanisms of synchronized gamma oscillations in inhibitory interneuron networks," *Nature reviews neuroscience*, vol. 8, no. 1, pp. 45–56, 2007.

[105] C. Zheng, K. W. Bieri, Y.-T. Hsiao, and L. L. Colgin, "Spatial sequence coding differs during slow and fast gamma rhythms in the hippocampus," *Neuron*, vol. 89, no. 2, pp. 398–408, 2016.

[106] K. W. Bieri, K. N. Bobbitt, and L. L. Colgin, "Slow and fast gamma rhythms coordinate different spatial coding modes in hippocampal place cells," *Neuron*, vol. 82, no. 3, pp. 670–681, 2014.

[107] A. B. Tort, R. W. Komorowski, J. R. Manns, N. J. Kopell, and H. Eichenbaum, "Theta–gamma coupling increases during the learning of item–context associations," *Proceedings of the National Academy of Sciences*, vol. 106, no. 49, pp. 20942–20947, 2009.

[108] P. R. Shirvalkar, P. R. Rapp, and M. L. Shapiro, "Bidirectional changes to hippocampal theta–gamma comodulation predict memory for recent spatial episodes," *Proceedings of the National Academy of Sciences*, vol. 107, no. 15, pp. 7054–7059, 2010.

[109] G. Buzsáki, D. L. Buhl, K. D. Harris, J. Csicsvari, B. Czéh, and A. Morozov, "Hippocampal network patterns of activity in the mouse," *Neuroscience*, vol. 116, no. 1, pp. 201–211, 2003.

[110] A. Bragin, C. L. Wilson, J. Almajano, I. Mody, and J. Engel Jr, "High-frequency oscillations after status epilepticus: epileptogenesis and seizure genesis," *Epilepsia*, vol. 45, no. 9, pp. 1017–1023, 2004.

[111] A. Ylinen, A. Bragin, Z. Nádasdy, G. Jandó, I. Szabo, A. Sik, and G. Buzsáki, "Sharp wave-associated high-frequency oscillation (200 hz) in the intact hippocampus: network and intracellular mechanisms," *Journal of Neuroscience*, vol. 15, no. 1, pp. 30–46, 1995.

[112] J. Engel Jr, A. Bragin, R. Staba, and I. Mody, "High-frequency oscillations: what is normal and what is not?," *Epilepsia*, vol. 50, no. 4, pp. 598–604, 2009.

[113] S. Burnos, P. Hilfiker, O. Sürücü, F. Scholkmann, N. Krayenbühl, T. Grunwald, and J. Sarnthein, "Human intracranial high frequency oscillations (hfos) detected by automatic time-frequency analysis," *PloS one*, vol. 9, no. 4, 2014.

[114] A. Bragin, G. Jando, Z. Nadasdy, M. Van Landeghem, and G. Buzsáki, "Dentate eeg spikes and associated interneuronal population bursts in the hippocampal hilar region of the rat," *Journal of neurophysiology*, vol. 73, no. 4, pp. 1691–1705, 1995.

[115] S. Lensu, T. Waselius, M. Penttonen, and M. S. Nokia, "Dentate spikes and learning: disrupting hippocampal function during memory consolidation can improve pattern separation," *Journal of neurophysiology*, vol. 121, no. 1, pp. 131–139, 2019.

[116] O. Devinsky, A. D. Patel, J. H. Cross, V. Villanueva, E. C. Wirrell, M. Privitera, S. M. Greenwood, C. Roberts, D. Checketts, K. E. VanLandingham, *et al.*, "Effect of cannabidiol on drop seizures in the lennox–gastaut syndrome," *New England Journal of Medicine*, vol. 378, no. 20, pp. 1888–1897, 2018.

[117] A. Pitkänen, W. Löscher, A. Vezzani, A. J. Becker, M. Simonato, K. Lukasiuk, O. Gröhn, J. P. Bankstahl, A. Friedman, E. Aronica, *et al.*, "Advances in the development of biomarkers for epilepsy," *The Lancet Neurology*, vol. 15, no. 8, pp. 843–856, 2016.

[118] W. Löscher, "The holy grail of epilepsy prevention: Preclinical approaches to antiepileptogenic treatments," *Neuropharmacology*, vol. 167, p. 107605, 2019.

[119] J. Engel Jr and A. Pitkänen, "Biomarkers for epileptogenesis and its treatment," *Neuropharmacology*, vol. 167, p. 107735, 2020.

[120] J. Engel Jr, "Epileptogenesis, traumatic brain injury, and biomarkers," *Neurobiology of disease*, vol. 123, pp. 3–7, 2019.

[121] J. Nissinen, P. Andrade, T. Natunen, M. Hiltunen, T. Malm, K. Kanninen, J. I. Soares, O. Shatillo, J. Sallinen, X. E. Ndode-Ekane, *et al.*, "Disease-modifying effect of atipamezole in a model of post-traumatic epilepsy," *Epilepsy research*, vol. 136, pp. 18–34, 2017.

[122] M. T. Kendirli, D. T. Rose, and E. H. Bertram, "A model of posttraumatic epilepsy after penetrating brain injuries: effect of lesion size and metal fragments," *Epilepsia*, vol. 55, no. 12, pp. 1969–1977, 2014.

[123] W. Turski, E. Cavalheiro, L. Calderazzo-Filho, Z. Kleinrok, S. Czuczwar, and L. Turski, "Injections of picrotoxin and bicuculline into the amygdaloid complex of the rat: an electroencephalographic, behavioural and morphological analysis," *Neuroscience*, vol. 14, no. 1, pp. 37–53, 1985.

[124] W. A. Turski, E. A. Cavalheiro, M. Schwarz, S. J. Czuczwar, Z. Kleinrok, and L. Turski, "Limbic seizures produced by pilocarpine in rats: behavioural, electroencephalographic and neuropathological study," *Behavioural brain research*, vol. 9, no. 3, pp. 315–335, 1983.

[125] L. J. Willmore, G. W. Sypert, and J. B. Munson, "Recurrent seizures induced by cortical iron injection: a model of posttraumatic epilepsy," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 4, no. 4, pp. 329–336, 1978.

[126] R. Schwarcz, R. Zaczek, and J. T. Coyle, "Microinjection of kainic acid into the rat hippocampus," *European Journal of pharmacology*, vol. 50, no. 3, pp. 209–220, 1978.

[127] M. R. Gluck, E. Jayatilleke, S. Shaw, A. J. Rowan, and V. Haroutunian, "Cns oxidative stress associated with the kainic acid rodent model of experimental epilepsy," *Epilepsy research*, vol. 39, no. 1, pp. 63–71, 2000.

[128] J. McNamara, "Kindling model of epilepsy.," *Advances in neurology*, vol. 44, pp. 303–318, 1986.

[129] J. J. Knierim, "The hippocampus," *Current Biology*, vol. 25, no. 23, pp. R1116–R1121, 2015.

[130] W. Deng, J. B. Aimone, and F. H. Gage, "New neurons and new memories: how does adult hippocampal neurogenesis affect learning and memory?," *Nature reviews neuroscience*, vol. 11, no. 5, pp. 339–350, 2010.

[131] D. Z. Milikovsky, I. Weissberg, L. Kamintsky, K. Lippmann, O. Schefenbauer, F. Frigerio, M. Rizzi, L. Sheintuch, D. Zelig, J. Ofer, *et al.*, "Electrocorticographic dynamics as a novel biomarker in five models of epileptogenesis," *Journal of Neuroscience*, vol. 37, no. 17, pp. 4450–4461, 2017.

[132] F. Vecchio, F. Miraglia, C. Vollono, F. Fuggetta, P. Bramanti, B. Cioni, and P. M. Rossini, "Pre-seizure architecture of the local connections of the epileptic focus examined via graph-theory," *Clinical Neurophysiology*, vol. 127, no. 10, pp. 3252–3258, 2016.

[133] M. Amiri, B. Frauscher, and J. Gotman, "Phase-amplitude coupling is elevated in deep sleep and in the onset zone of focal epileptic seizures," *Frontiers in human neuroscience*, vol. 10, p. 387, 2016.

[134] I. A. Nissen, C. J. Stam, E. C. van Straaten, V. Wottschel, J. C. Reijneveld, J. C. Baayen, P. C. de Witt Hamer, S. Idema, D. N. Velis, and A. Hillebrand, "Localization of the epileptogenic zone using interictal meg and machine learning in a large cohort of drug-resistant epilepsy patients," *Frontiers in neurology*, vol. 9, p. 647, 2018.

[135] H.-J. Huppertz, E. Hof, J. Klisch, M. Wagner, C. H. Lücking, and R. Kristeva-Feige, "Localization of interictal delta and epileptiform eeg activity associated with focal epileptogenic brain lesions," *Neuroimage*, vol. 13, no. 1, pp. 15–28, 2001.

[136] J. S. Naftulin, O. J. Ahmed, G. Piantoni, J.-B. Eichenlaub, L.-E. Martinet, M. A. Kramer, and S. S. Cash, "Ictal and preictal power changes outside of the seizure focus correlate with seizure generalization," *Epilepsia*, vol. 59, no. 7, pp. 1398–1409, 2018.

[137] C. Cuello-Oderiz, N. von Ellenrieder, F. Dubeau, and J. Gotman, "Influence of the location and type of epileptogenic lesion on scalp interictal epileptiform discharges and high-frequency oscillations," *Epilepsia*, vol. 58, no. 12, pp. 2153–2163, 2017.

[138] L. Li, M. Patel, J. Almajano, J. Engel Jr, and A. Bragin, "Extrahippocampal high-frequency oscillations during epileptogenesis," *Epilepsia*, vol. 59, no. 4, pp. e51–e55, 2018.

[139] P. Andrade, J. Nissinen, and A. Pitkänen, "Generalized seizures after experimental traumatic brain injury occur at the transition from slow-wave to rapid eye movement sleep," *Journal of neurotrauma*, vol. 34, no. 7, pp. 1482–1487, 2017.

[140] L. Sheybani, G. Birot, A. Contestabile, M. Seeck, J. Z. Kiss, K. Schaller, C. M. Michel, and C. Quairiaux, "Electrophysiological evidence for the development of a self-sustained large-scale epileptic network in the kainate mouse model of temporal lobe epilepsy," *Journal of Neuroscience*, vol. 38, no. 15, pp. 3776–3791, 2018.

[141] M. Rizzi, C. Brandt, I. Weissberg, D. Z. Milikovsky, A. Pauletti, G. Terrone, A. Salamone, F. Frigerio, W. Löscher, A. Friedman, *et al.*, "Changes of dimension of EEG/ECoG nonlinear dynamics predict epileptogenesis and therapy outcomes," *Neurobiology of disease*, vol. 124, pp. 373–378, 2019.

[142] M. L. Goodenberger and R. B. Jenkins, "Genetics of adult glioma," *Cancer genetics*, vol. 205, no. 12, pp. 613–621, 2012.

[143] E. Hattingen, P. Raab, K. Franz, H. Lanfermann, M. Setzer, R. Gerlach, F. E. Zanella, and U. Pilatus, "Prognostic value of choline and creatine in who grade ii gliomas," *Neuroradiology*, vol. 50, no. 9, pp. 759–767, 2008.

[144] R. J. Gillies and D. L. Morse, "In vivo magnetic resonance spectroscopy in cancer," *Annu. Rev. Biomed. Eng.*, vol. 7, pp. 287–326, 2005.

[145] S. K. Gujar, S. Maheshwari, I. Björkman-Burtscher, and P. C. Sundgren, "Magnetic resonance spectroscopy," *Journal of neuro-ophthalmology*, vol. 25, no. 3, pp. 217–226, 2005.

[146] M. E. Watts, R. Pocock, and C. Claudianos, "Brain energy and oxygen metabolism: emerging role in normal function and disease," *Frontiers in molecular neuroscience*, vol. 11, p. 216, 2018.

[147] D. Soares and M. Law, "Magnetic resonance spectroscopy of the brain: review of metabolites and clinical applications," *Clinical radiology*, vol. 64, no. 1, pp. 12–21, 2009.

[148] E. Hattingen, P. Raab, K. Franz, F. E. Zanella, H. Lanfermann, and U. Pilatus, "Myo-inositol: a marker of reactive astrogliosis in glial tumors?," *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo*, vol. 21, no. 3, pp. 233–241, 2008.

[149] N. Sibtain, F. Howe, and D. Saunders, "The clinical value of proton magnetic resonance spectroscopy in adult brain tumours," *Clinical radiology*, vol. 62, no. 2, pp. 109–119, 2007.

[150] E. Kousi, I. Tsougos, and K. Eftychia, "Proton magnetic resonance spectroscopy of the central nervous system," *Novel Frontiers of Advanced Neuroimaging. InTech*, vol. 2013, pp. 19–50, 2013.

[151] R. A. De Graaf, *In vivo NMR spectroscopy: principles and techniques*. John Wiley & Sons, 2019.

[152] D. Yang, Y. Korogi, T. Sugahara, M. Kitajima, Y. Shigematsu, L. Liang, Y. Ushio, and M. Takahashi, "Cerebral gliomas: prospective comparison of multivoxel 2d chemical-shift imaging proton mr spectroscopy, echoplanar perfusion and diffusion-weighted mri," *Neuroradiology*, vol. 44, no. 8, pp. 656–666, 2002.

[153] A. A. Tzika, L. G. Astrakas, M. K. Zarifi, D. Zurakowski, T. Y. Poussaint, L. Goumnerova, N. J. Tarbell, and P. M. Black, "Spectroscopic and perfusion magnetic resonance imaging predictors of progression in pediatric brain tumors," *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 100, no. 6, pp. 1246–1256, 2004.

[154] E. R. Danielsen and B. Ross, *Magnetic resonance spectroscopy diagnosis of neurological diseases*. CRC Press, 1999.

[155] F. Howe, S. Barton, S. Cudlip, M. Stubbs, D. Saunders, M. Murphy, P. Wilkins, K. Opstad, V. Doyle, M. McLean, *et al.*, "Metabolic profiles of human brain tumors using quantitative in vivo 1h magnetic resonance spectroscopy," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 49, no. 2, pp. 223–232, 2003.

[156] S. K. Natarajan and S. Venneti, "Glutamine metabolism in brain tumors," *Cancers*, vol. 11, no. 11, p. 1628, 2019.

[157] S. Ishiuchi, K. Tsuzuki, Y. Yoshida, N. Yamada, N. Hagimura, H. Okado, A. Miwa, H. Kurihara, Y. Nakazato, M. Tamura, *et al.*, "Blockage of ca 2+-permeable ampa receptors suppresses migration and induces apoptosis in human glioblastoma cells," *Nature medicine*, vol. 8, no. 9, pp. 971–978, 2002.

[158] I. Mader, S. Rauer, P. Gall, and U. Klose, "1h mr spectroscopy of inflammation, infection and ischemia of the brain," *European journal of radiology*, vol. 67, no. 2, pp. 250–257, 2008.

[159] G. Fan, "Magnetic resonance spectroscopy and gliomas," *Cancer Imaging*, vol. 6, no. 1, pp. 113–115, 2006.

[160] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[161] P. D. Emmady and A. C. Anilkumar, *EEG Abnormal Waveforms*. StatPearls Publishing, Treasure Island (FL), 2020.

[162] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 185–194, 2019.

[163] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, "Long-term feature banks for detailed video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 284–293, 2019.

[164] P. Sudharshan, C. Petitjean, F. Spanhol, L. E. Oliveira, L. Heutte, and P. Honeine, "Multiple instance learning for histopathological breast cancer image classification," *Expert Systems with Applications*, vol. 117, pp. 103–111, 2019.

[165] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah, "Activation atlas," *Distill*, 2019. https://distill.pub/2019/activation-atlas.

[166] G. Pruthi, F. Liu, S. Kale, and M. Sundararajan, "Estimating training data influence by tracing gradient descent," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[167] L. Smyth, D. Kangin, and N. Pugeault, "Training-valuenet: Data driven label noise cleaning on weakly-supervised web images," in *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 307–312, IEEE, 2019.

[168] B. R. Cowley and J. W. Pillow, "High-contrast" gaudy" images improve the training of deep neural network models of visual cortex," *arXiv preprint arXiv:2006.11412*, 2020.

[169] G. Pleiss, T. Zhang, E. R. Elenberg, and K. Q. Weinberger, "Identifying mislabeled data using the area under the margin ranking," *arXiv preprint arXiv:2001.10528*, 2020.

[170] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[171] N. Olliverre, G. Yang, G. Slabaugh, C. C. Reyes-Aldasoro, and E. Alonso, "Generating magnetic resonance spectroscopy imaging data of brain tumours from linear, non-linear and deep learning models," in *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 130–138, Springer, 2018.