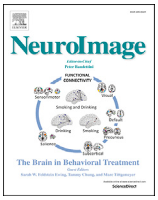
NeuroImage 220 (2020) 117021

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



Machine-learning-based diagnostics of EEG pathology

Lukas A.W. Gemein [a,b,c,*], Robin T. Schirrmeister [a,b], Patryk Chrabąszcz [a,b], Daniel Wilson [a], Joschka Boedecker [c], Andreas Schulze-Bonhage [d], Frank Hutter [b], Tonio Ball [a,d]

[a] Neuromedical AI Lab, Department of Neurosurgery, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Engelbergerstr. 21, 79106, Freiburg, Germany

[b] Machine Learning Lab, Computer Science Department – University of Freiburg, Faculty of Engineering, University of Freiburg, Georges-Köhler-Allee 74, 79110, Freiburg, Germany

[c] Neurorobotics Lab, Computer Science Department – University of Freiburg, Faculty of Engineering, University of Freiburg, Georges-Köhler-Allee 80, 79110, Freiburg, Germany

[d] Freiburg Epilepsy Center, Department of Neurosurgery, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Breisacher Str. 64, 79106, Freiburg, Germany

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Electroencephalography
EEG
Diagnostics
Pathology
Features
Riemannian geometry
Convolutional neural networks

ABSTRACT

Machine learning (ML) methods have the potential to automate clinical EEG analysis. They can be categorized into feature-based (with handcrafted features), and end-to-end approaches (with learned features). Previous studies on EEG pathology decoding have typically analyzed a limited number of features, decoders, or both. For a I) more elaborate feature-based EEG analysis, and II) in-depth comparisons of both approaches, here we first develop a comprehensive feature-based framework, and then compare this framework to state-of-the-art end-to-end methods. To this aim, we apply the proposed feature-based framework and deep neural networks including an EEG-optimized temporal convolutional network (TCN) to the task of pathological versus non-pathological EEG classification. For a robust comparison, we chose the Temple University Hospital (TUH) Abnormal EEG Corpus (v2.0.0), which contains approximately 3000 EEG recordings. The results demonstrate that the proposed feature-based decoding framework can achieve accuracies on the same level as state-of-the-art deep neural networks. We find accuracies across both approaches in an astonishingly narrow range from 81 to 86%. Moreover, visualizations and analyses indicated that both approaches used similar aspects of the data, e.g., delta and theta band power at temporal electrode locations. We argue that the accuracies of current binary EEG pathology decoders could saturate near 90% due to the imperfect inter-rater agreement of the clinical labels, and that such decoders are already clinically useful, such as in areas where clinical EEG experts are rare. We make the proposed feature-based framework available open source and thus offer a new tool for EEG machine learning research.

1. Introduction

There is a great interest in using machine learning (ML) methods for automatic electroencephalogram (EEG) analysis, especially in the domain of clinical diagnostics based on the EEG. For example, it forms a basis for detecting and predicting epileptic seizures [Subasi et al., 2019; Hügle et al., 2018; Kiral-Kornek et al., 2018; Mirowski et al., 2009] with the goal of warning patients of upcoming seizures or to control brain stimulation for preventing or stopping seizure activity. Furthermore, ML allows for the automation of the process of EEG-based sleep staging [Biswal et al. (2017)] and neurological diagnostics of both specific diseases and disorders such as Alzheimer's disease [Lehmann et al. (2007)], depression [Cai et al. (2016); Hosseinifard et al. (2013)], traumatic brain injuries [Albert et al. (2016)], strokes [Giri et al. (2016)], and disorders of consciousness [Engemann et al. (2018); Sun et al. (2019)], or of general EEG pathology [Lopez de Diego (2017); Schirrmeister et al. (2017a); Roy et al. (2019a); Amin et al. (2019); Alhussein et al. (2019); Van Leeuwen et al. (2019)].

There are several facts that motivate the interest in automatic clinical EEG diagnosis. First, the evaluation of clinical EEGs is frequently a time-consuming and exhausting process. Second, it requires years of training to assess pathological changes in clinical EEG recordings. Moreover, even for highly trained EEG experts, diagnostic accuracy is subject to a number of limitations. It depends highly on individual training and experience,

* Corresponding author. Neuromedical AI Lab, Department of Neurosurgery, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Engelbergerstr. 21, 79106, Freiburg, Germany.
E-mail address: lukas.gemein@uniklinik-freiburg.de (L.A.W. Gemein).

https://doi.org/10.1016/j.neuroimage.2020.117021
Received 17 December 2019; Received in revised form 16 March 2020; Accepted 26 May 2020
Available online 10 June 2020

consistency of rating over time, time constraints in different filter settings of frequently subjectively defined frequency bands, and unclear criteria for the thresholding of potential changes, e.g., at low amplitude in relation to the background EEG. Accordingly, inter-rater agreement in assessing EEGs is known to be moderate [Landis and Koch (1977)], i.e., Grant et al. (2014) found a Fleiss' Kappa of 0.44 when neurologists classified recordings to one of seven classes including seizure, slowing, and normal activity. In the more general task of classifying EEG recordings as pathological or normal, Houfek and Ellingson (1959) and Rose et al. (1973) reported inter-rater agreements of 86% and 88% based on two neurologists. The development of algorithms for automated EEG diagnostics could support clinicians in screening EEGs. They could not only reduce the workload of clinicians, but also allow for earlier detection and treatment of diseases, which could enhance patient care. Furthermore, they could provide high-quality EEG interpretation and classification to patients that cannot attend specialized centers.

We broadly categorize ML for EEG analysis into two approaches: feature-based and end-to-end methods. Feature-based decoding methods have a long history of successful application in different EEG decoding tasks. In this approach, typically handcrafted and *a priori* selected features represent the data. For example, a researcher could *a priori* decide to use the spectral power in certain frequency bands as features, if they assume that these bands are informative for the decoding task at hand. The choice of exact frequency bands could then be handcrafted, such as in the common spatial patterns (CSP) algorithm for motor decoding [Müller-Gerking et al. (1999)], or they could be determined by automatic feature selection, such as by the recursive band elimination in the filter bank CSP (FBCSP) algorithm [Ang et al. (2008)]. This procedure relies on the domain expertise of the researcher. If the *a priori* feature decisions are sub-optimal, it can diminish the quality of the resulting analysis. Conversely, owing to its explicit nature, interpretability of the classification decisions is frequently considered an advantage of feature-based decoding.

Conversely, end-to-end decoding methods accept raw or minimally preprocessed data as inputs. To date, end-to-end deep learning has attracted attention primarily owing to its success in other research fields, such as computer vision [Krizhevsky et al. (2012)] and speech recognition [Hinton et al. (2012)]. However, it has also recently gained momentum through the successful application of deep learning with artificial neural networks to EEG analysis [Roy et al. (2019b); Craik et al. (2019)]. By design, the networks learn features themselves and allow for a joint optimization of the feature extraction and classification. This procedure can lead to superior solutions or the discovery of unexpected informative features and does not require handcrafting, at least not for the extraction of the features. End-to-end models have the reputation for being "black boxes" with regard to the learned features; it is a challenge and an ongoing topic of intense research in the deep learning field to understand what they have learned [Montavon et al. (2018); Sturm et al. (2016); Hartmann et al. (2018)]. Another common concern is that the complexity in the application of ML is only shifted from the domain of feature engineering in traditional approaches to the domain of network engineering inasmuch as it could be necessary to handcraft the networks according to the requirements of a given task.

In the literature, there is a lack of systematic comparisons of traditional feature-based versus end-to-end ML analysis of EEGs, despite their importance for a wide range of applications. In particular, there are no studies comparing the accuracy of pathology decoding from EEGs using a broad range of time, frequency, and connectivity features with well-established end-to-end methods based on a large EEG data set for a robust comparison. Past comparisons of deep learning results have frequently only considered other deep learning results, or (rather) simple feature-based baselines (using thresholds, linear discriminant analysis, or linear regression) with limited feature sets. This can lead to unfair comparisons among the methods. Moreover, it can create the impression of superiority of one approach over another. It is actually possible that deep learning could not yield an improvement over feature-based decoding in specific applications. Recently, Rajkomar et al. (2018) demonstrated that logistic regression can compete with deep neural networks in predicting medical events from electronic health records. To the best of our knowledge, there is no work that compares different deep neural network architectures with a feature-based approach, especially using a large set of features of several domains to decode the EEGs. However, we anticipate that large-scale comparisons between feature-based and end-to-end methods will be critical to the advancement of ML techniques for EEGs beyond the current state-of-the-art. Developing methods in both of these important fields in a mutually informed manner will likely be fruitful for both advanced feature-based and novel end-to-end EEG methodologies.

Table 1
Related works on pathology decoding using TUH Abnormal EEG Corpus. All approaches rely on ConvNet architectures. Only chronologically oldest publication used handcrafted features. Publications marked with * used pretrained models and additional training data. Publication marked with $^+$ did not use TUH Abnormal EEG Corpus.

Automated EEG Diagnosis	Features	Architecture	ACC [%]
Lopez de Diego (2017)	Cepstral coeff.	CNN + MLP	78.8
Schirrmeister et al. (2017a)		BD-Deep4	85.4
Roy et al. (2019a)		ChronoNet	86.6
Amin et al. (2019)*		AlexNet + SVM	87.3
Alhussein et al. (2019)*		3 x AlexNet + MLP	89.1
Van Leeuwen et al. (2019)$^+$		BD-Deep4	82.0

In this paper, we compare end-to-end decoding using deep neural networks to feature-based decoding using a large set of features. We design a comprehensive study using the Temple University Hospital (TUH) Abnormal EEG Corpus [Lopez de Diego (2017)] with approximately 3000 recordings of at least 15 min duration each. This is a subset of the TUH EEG Corpus [Obeid and Picone (2016)], the largest publicly available collection of EEG recordings to date. For feature-based pathology decoding, we use random forest (RF) [Breiman (2001)], support vector machine (SVM) [Boser et al. (1992)], Riemannian geometry (RG), and the auto-sklearn calssifier (ASC) [Feurer et al. (2015)] – an automated ML toolkit. For end-to-end pathology decoding, we use three types of convolutional neural networks (ConvNets, in other publications also CNN) [LeCun et al. (1999)] that have a history of successful application in different EEG decoding tasks. These are the 4-layer ConvNet architecture Braindecode Deep4 ConvNet (BD-Deep4), which has been successfully applied to motor decoding [Schirrmeister et al. (2017b)], velocity and speed decoding [Hammer et al. (2013)], and pathology decoding [Schirrmeister et al. (2017a); Van Leeuwen et al. (2019)]. Importantly, we use Braindecode (BD),[1] a previously developed and evaluated deep learning toolbox for EEGs, "out of the box" – without task-specific network engineering, i.e., without adaptation to the architectures. Furthermore, we use a TCN [Bai et al. (2018)] that is optimized for EEG decoding with a neural architecture search. We call this adaptation BD-TCN.

To the best of our knowledge, there are currently six published results for pathology decoding from EEGs, five of which used the TUH Abnormal EEG Corpus (Table 1). However, only one publication [Lopez de Diego (2017)] used handcrafted features and a classification through a CNN with a multi layer perceptron (MLP). All other papers considered this initial feature-based decoding result as a baseline. Whereas Amin et al. (2019) and Alhussein et al. (2019) have reported the highest accuracies in decoding pathology from EEGs, we exclude them from our direct comparison. The papers mention "pretrained models" and additional "10,000 normal EEG recordings", which would appear to be an extension of the TUH Abnormal EEG data set without specifying more details. In ML, the effect of more data is commonly greater than the effect of more elaborate algorithms [Halevy et al. (2009)]. A direct comparison of

[1] Available for download at https://github.com/TNTLFreiburg/braindecode.

Table 2
Number of recordings and patients in TUH Abnormal EEG Corpus (v2.0.0). For certain patients, there exist several recordings. For other patients in the development set, there exist normal and abnormal recordings. There is no overlap of patients in development and the final evaluation set.

TUH Abnormal EEG	Non-pathological		Pathological		Intersection
Corpus (v2.0.0)	Recordings	Patients	Recordings	Patients	Patients
Development set	1371	1237	1346	893	54
Final evaluation set	150	148	126	105	0
Total	1521	1385	1472	998	54

publications with access to a substantially larger amount of training data would hence be unfair.

The paper is structured as follows. In Section 2, we provide an introduction to the TUH Abnormal EEG Corpus upon which we base our study. We then discuss the feature and deep learning pipeline in detail and explain how we proceeded in evaluating and comparing both approaches. The section closes with a discussion of the analytical methods we used to assist in our interpretation of the results. In Section 3, we present and discuss our results including an extensive comparison of both pipelines. We present a general discussion in Section 4, and close with a brief outlook and conclusions in Section 5.

2. Material and methods

2.1. Data

We base our study on the TUH Abnormal EEG Corpus[2] (v2.0.0), which is currently unique owing to its size and public availability and has enabled the task of general pathology decoding from EEGs. The corpus includes 2993 recordings of at least 15 min duration obtained from 2329 unique patients and consists of a development and separate final evaluation set (Table 2). It contains recordings of both male and female patients of a wide age range (7 days–96 years), thus including infants, children, adolescents, adults, and senior patients. Pathologies diagnosed in the patients in the data set include (but are not limited to) epilepsy, strokes, depression, and Alzheimer's disease, however, only binary labels are provided. The data set includes physician reports that provide additional information regarding each EEG recording, such as main EEG findings, ongoing medication of the patient, and medical history. In the description of the data set,[3] the TUH reports an inter-rater agreement of 97–100%. In the literature, the reported scores are typically considerably lower [Houfek and Ellingson (1959); Rose et al. (1973)]. The almost perfect rating scores could be a consequence of the review process of the findings that were performed by medical students that knew the diagnoses beforehand [Picone (2019)]. For more information on the data set see Lopez de Diego (2017) and Obeid and Picone (2016).

2.2. Common preprocessing in both feature and end-to-end pipeline

Typically, at least minimal preprocessing of the raw EEG data is applied in both scenarios, relying on handcrafted feature extraction and based on end-to-end approaches. We applied the preprocessing steps described below to both scenarios to normalize the input distribution and thus stabilize the deep network learning process, a common practice in deep learning applications, and to stabilize feature extraction. However, the latter requires additional steps that are described in Section 2.4. Importantly, our general preprocessing did not preselect any EEG features. As in our earlier work on EEG pathology decoding with deep ConvNets [Schirrmeister et al. (2017a)], we included the following preprocessing steps: First, we selected a subset of 21 electrode positions (Fig. 1) following the international 10–20 placement [Jasper (1958)] because these electrode positions occurred in all the individual recordings in the data set. Then, we discarded the first 60 s of every recording because we observed a large number of recording artifacts in this period, which could have been caused by rearrangement of the electrode cap or by finding a comfortable seating position. Moreover, we used a maximum of 20 min of every recording to avoid considerable feature generation and resampling times for exceptionally long recordings. As in our previous work Schirrmeister et al. (2017a) and in the work by Van Leeuwen et al. (2019), EEG recordings were downsampled to 100 Hz and clipped at $\pm 800 \mu V$ to reject unphysiologically extreme values and to ensure comparability to these previous studies. Although Roy et al. (2019a) performed their experiments at 250 Hz, we chose to use 100 Hz for better comparability with the other approaches, and to avoid motor artifacts. However, this could place us at a disadvantage in the direct comparison with Roy et al. (2019a). Our preprocessing partially uses code from Python libraries MNE[4] and resampy.[5]

2.3. End-to-end decoding with deep neural networks

2.3.1. Neural network architectures

We used different neural network architectures including ConvNets and TCNs to decode the pathology from the EEG recordings. First, we used a four-layered ConvNet architecture called BD-Deep4 as previously

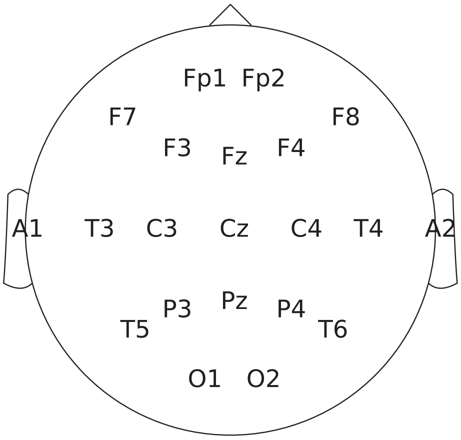


Fig. 1. Topographical map of 21 EEG channel subset of international 10–20 placement [Jasper (1958)] common in all recordings included in TUH Abnormal EEG Corpus (v2.0.0).

[2] Available for download at https://www.isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml.

[3] Available for download at https://www.isip.piconepress.com/projects/tuh_eeg/downloads/tuh_eeg_abnormal/v2.0.0/_AAREADME.txt.

[4] Available for download at https://github.com/mne-tools/mne-python.

[5] Available for download at https://github.com/bmcfee/resampy.

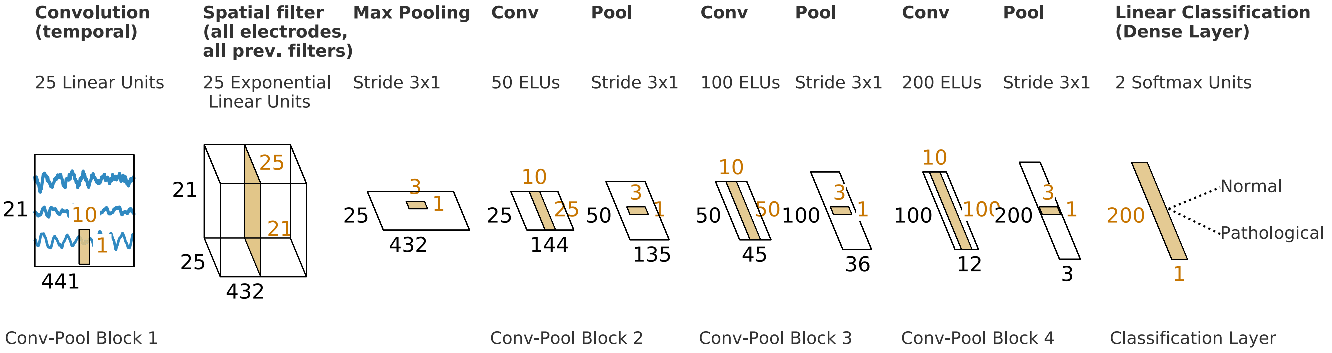


Fig. 2. Four-layered BD-Deep4 as introduced by Schirrmeister et al. (2017b). Initial separated convolution is followed by several convolution and max-pooling blocks.

introduced by Schirrmeister et al. (2017b). The BD-Deep4 architecture [Fig. 2] has an initial separated convolution[6] (first temporal, then spatial). Subsequently, it has several blocks consisting of convolution and max-pooling and uses exponential linear units as activation functions. It is a rather general architecture that has proven to generalize well to several EEG decoding tasks such as motor (imagery) decoding [Schirrmeister et al. (2017b)], velocity and speed decoding [Hammer et al. (2013)], and pathology decoding [Schirrmeister et al. (2017a); Van Leeuwen et al. (2019)]. We applied BD-Deep4 without any further adjustments to its architecture.

Next, we used a TCN architecture [Fig. 3] that was evaluated in a Master thesis by Chrabąszcz (2018). The TCN was originally proposed by Bai et al. (2018) as an alternative to recurrent neural networks (RNN) [Rumelhart et al. (1988)]. In their work, Bai et al. (2018) demonstrate that the TCN consistently outperforms RNNs in sequence modeling tasks across different datasets that are commonly used for benchmarking RNNs. It is the most complex and deepest architecture under investigation in the present study. The optimization by Chrabąszcz (2018) resulted in five levels of blocks consisting of temporal convolutions with 55 channels each as well as max-pooling. We call this optimized architecture Braindecode TCN (BD-TCN). For more information on the optimized hyperparameters, see Table S2.

Furthermore, we used another ConvNet architecture introduced by Schirrmeister et al. (2017b) called Braindeocde Shallow ConvNet (BD-Shallow). The network [Fig. 4], as in the BD-Deep4 network, has an initial separated convolution; however, it is the only convolution in the entire architecture. The well-known FBCSP algorithm [Ang et al. (2008)] inspired the BD-Shallow architecture, in particular the squaring and logarithmic nonlinearities. It was designed to specifically extract the logarithm of the band power of EEG signals. We applied BD-Shallow, as BD-Deep, without any further adjustments to its architecture.

Moreover, we used a reimplementation of another ConvNet architecture called EEGNet that was originally introduced by Lawhern et al. (2018). We call this reimplementation Braindecode EEGNet (BD-EEGNet). Again, the architecture has a separated initial convolution. Furthermore, the architecture is remarkable owing to its small number of parameters (see Table S3).

2.3.2. *Training of neural networks*

We trained the networks in a cropped manner with equally-sized, maximally overlapping crops as described by Schirrmeister et al. (2017b). The receptive field of the networks automatically determines the size of the crops. All networks are exposed to approximately 600 signal samples at a time, except the TCN which has a receptive field of approximately 900 samples [Table S3]. Unlike the original paper by Schirrmeister et al. (2017a), we used optimizer AdamW [Loshchilov and Hutter (2017)] over Adam [Kingma and Ba (2014)] to minimize the categorical cross-entropy loss function. AdamW decouples weight decay updates and the optimization of the loss function, which allows for better generalization [Loshchilov and Hutter (2017)]. We used cosine annealing [Loshchilov and Hutter (2016)] to schedule the learning rates for the gradient and weight decay updates. We did not perform learning rate restarts.

2.4. *Feature-based decoding*

2.4.1. *Additional preprocessing prior to feature extraction*

After general preprocessing common to both pipelines (see Section 2.2), we applied several additional steps in the feature-based pipeline. In the special case of connectivity feature extraction, we first filtered entire signals to a selected frequency range in the time domain to avoid the creation of filtering artifacts at the start and end points of the signal segments. We split every recording into equally sized, nonoverlapping signal segments called crops of 600 samples, i.e., given the sampling frequency of 100 Hz, this corresponded to 6 s, to be maximally comparable to the end-to-end pipeline, where the receptive field of the architectures determines the crop size [S3]. We discarded crops with values of $\pm 800 \mu V$ to stabilize feature generation. This resulted in the exclusion of one recording (subject 00008184, session s001_t001), as one channel was not properly recorded, meaning each measurement at every time point exceeded the outlier value.

2.4.2. *Feature extraction*

Our proposed concept of feature-based decoding shows significant differences to those presented in the literature. Where we computed 8633 features of 50 feature types and six domains [see Table 3], typically far smaller feature sets were used in the literature. For example, Hosseinifard et al. (2013) extracted, exclusively, the total power from the theta, alpha, and beta frequency bands to decode depression. Similarly, Lopez de Diego (2017) extracted a single feature type (cepstral coefficients) that originated from the field of speech recognition to decode EEG pathology. An example of a study with a larger amount of feature types is the work of Cai et al. (2016). They extracted 16 feature types including amplitude, time, and connectivity measures to detect mild depression. In our present study, however, we computed an even larger feature set with more than twice the number of feature types [Table 3].

We computed a large set of features describing time, frequency, and connectivity structure of the EEG signals that have all been used to characterize EEGs [Subasi (2007), Logesparan et al. (2012), Kuhlmann et al. (2008), Kumar et al. (2010), Quiroga et al. (1997), Hjorth (1970), James and Lowe (2003), Petrosian (1995), Inouye et al. (1991), Roberts et al. (1999), Balli and Palaniappan (2009), Peng et al. (1995), Watter

[6] ”Separated convolution” is an overloaded term. For a detailed explanation of what is meant in the context of the ConvNets used for this study please refer to Section 2.4.1 in Schirrmeister et al. (2017b), where the models where originally introduced.

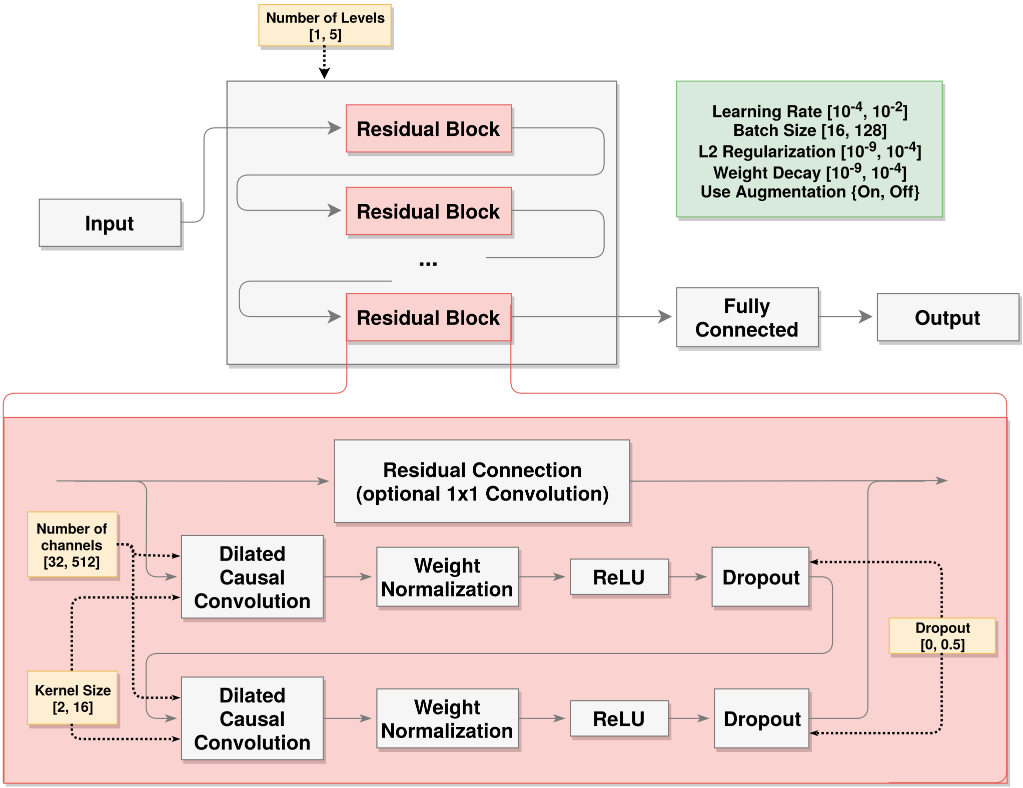

Number of Levels [1, 5]
Input
Residual Block
Residual Block
...
Residual Block
Learning Rate [10^-4, 10^-2]
Batch Size [16, 128]
L2 Regularization [10^-9, 10^-4]
Weight Decay [10^-9, 10^-4]
Use Augmentation {On, Off}
Fully Connected
Output
Residual Connection (optional 1x1 Convolution)
Number of channels [32, 512]
Kernel Size [2, 16]
Dilated Causal Convolution
Weight Normalization
ReLU
Dropout
Dropout [0, 0.5]
Dilated Causal Convolution
Weight Normalization
ReLU
Dropout

Fig. 3. General architecture of TCN as introduced by Bai et al. (2018) and search base explored in Master thesis by Chrabąszcz (2018) to find BD-TCN.

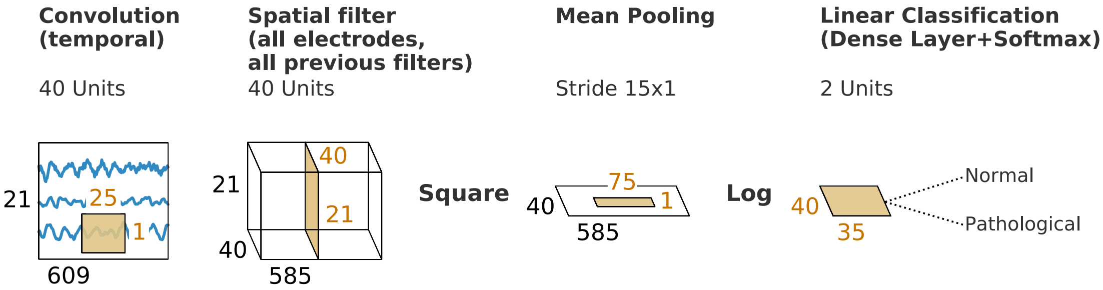


Fig. 4. BD-Shallow architecture as originally introduced by Schirrmeister et al. (2017b) inspired by FBCSP algorithm [Ang et al. (2008)].

(2014), Minasyan et al. (2010), van Putten et al. (2005), Esteller et al. (2001), Katz (1988), Lachaux et al. (1999)]. We generated features on every crop based on Discrete Fourier Transform (FT), Continuous and Discrete Wavelet Transform (CWT and DWT), and between-electrode connectivity features based on the Hilbert Transform [Table 3]. Furthermore, we parsed age and gender of the patients from the European Data Format [Kemp et al. (1992)] recording file headers as optional additional features. The feature implementations partially use code from the Python libraries PyEEG[7] and PyWavelets.[8] The implementations can be found in our feature decoding toolbox Brainfeatures.[9]

For the CWT, DWT, and FT feature computation, we weighted time domain crops with a Blackman-Harris window function to enhance the spectral estimation and reduce the effect of leakage. In preliminary experiments [Gemein (2017)], we tested different window functions. The Blackman-Harris window yielded best results, although the choice of the window function had only minor effects on the decoding accuracy. We extracted frequency features from bands 0–2 Hz, 2–4 Hz, 4–8 Hz, 8–13 Hz, 13–18 Hz, 18–24 Hz, 24–30 Hz, and 30–50 Hz using FT and a band overlap of 50%. We chose the bands to match the frequency bands commonly used in the literature. Furthermore, we observed superior results in preliminary experiments when using a band overlap [Gemein (2017)]. We chose wavelet scales for CWT and levels for DWT to match these bands as closely as possible.

For connectivity feature computation, we transformed the frequency-filtered time-crops using the Hilbert Transform to extract the signal phase.

The dimension of the feature vectors, including all feature values of domains CWT, DWT, FT, Connectivity, and Time, was $F = 8631$.

Time-resolved features Feature generation resulted in a feature matrix $M_i \in \mathbb{R}^{C_i \times F}$ for every recording, where I is the total number of recordings, C_i is the number of analyzed 6-s crops $i \in I$, and F is the dimension of the feature vector. For time-resolved (non-aggregated) decoding, we considered every feature vector of every crop as an independent example. This drastically increased the number of training examples, which could be beneficial in the training phase. However, it also resulted in higher memory consumption and higher learning times.

Aggregated features For aggregated decoding, we computed the

[7] Available for download at https://github.com/forrestbao/pyeeg.

[8] Available for download at https://github.com/PyWavelets/pywt.

[9] Available for download at https://github.com/TNTLFreiburg/brainfeatures.

Table 3
All implemented features sorted by feature domain. Feature domains are CWT/DWT, FT, Patient information, RG, Connectivity, and Time. Features marked with * were computed using PyEEG. Features marked with $^+$ were computed using PyWavelets.

CWT/DWT$^+$	**FT**	**Patient information**
Bounded variation	Maximum	Age
Maximum	Mean	Gender
Mean	Minimum	
Minimum	Peak frequency	**Riemannian**
Power	Power	Covariance matrix
Power ratio	Power ratio	
Spectral entropy	Spectral entropy	**Connectivity**
Variance	Value range	Phase Locking Value
	Variance	
Time	Hjorth mobility	Minimum
Detrended Fluctuation Analysis*	Hurst exponent	Nonlinear energy
Energy	Kurtosis	Petrosian fractal dimension*
Fisher information*	Line length	Skewness
Fractal dimension	Lyauponov exponent*	SVD entropy*
Higuchi fractal dimension	Maximum	Zero crossings
Hjorth activity	Mean	Zero crossings of derivative
Hjorth complexity	Median	

Table 4
Result of ensembling investigation indicating models, weights, and performances of two ensembles. One based on majority voting and one automatically selected by auto-sklearn. Neither ensemble improved compared to best single-model performances.

	Majority Vote		Auto-sklearn	
	Model	Weight	Model	Weight
	BD-Deep4	1	BD-Deep4	0.048
	RF	1	BD-Shallow	0.286
	RG	1	BD-TCN	0.190
			BD-EEGNet	0.238
			RF	0.190
			RG	0.048
CV [%]	84.61		86.23	
Final evaluation [%]	85.51		85.14	

aggregate of all time-crop feature matrices $M_i \in \mathbb{R}^{C_i \times F}$. Therefore, we used the median as the aggregation function, such that we obtained a single feature vector of length F for each recording. In previous experiments [Gemein (2017)], the median proved to be the best aggregation function in terms of decoding accuracy, although, again, the choice only had a minor effect. Aggregation drastically reduces the feature matrix size, which allows for faster learning and prediction. However, it has the disadvantage of discarding all time-resolved information as it collapses features of all crops of a recording into a single feature vector. The shape of the final aggregated feature matrix was $M_{aggregate} \in \mathbb{R}^{I \times F}$.

Dimensionality reduction We reduced the feature dimension F exclusively in preliminary experiments using principal component analysis (PCA) [Wold et al. (1987)]. Independent of the choice of principal components or ratio of variance, the application of PCA consistently led to a decrease in decoding accuracy.

Covariance matrices For Riemannian-geometry-based decoding (see below) we computed a covariance feature matrix $\Sigma_i \in \mathbb{R}^{C_i \times E^* E}$ for every crop, where E is the number of electrodes. Therefore, we used the Python package pyRiemann[10]. We independently tested the Euclidean and the geometric means to aggregate covariance matrices of the crops, such that we obtained a feature vector of length $E^*(E+1)/2$ for each recording and aggregation type. The shape of the final covariance feature matrix was $M_{riemann} \in \mathbb{R}^{I \times E^*(E+1)/2}$.

2.4.3. *Feature-based classifiers*

After feature generation, we used the feature matrix as an input to several feature-based ML models. We used an SVM with radial basis function (RBF) kernel, as commonly used in the literature [Lehmann et al. (2007); Cai et al. (2016)]. Furthermore, we used an RF classifier which is, by design, robust towards overfitting and hence a reliable baseline model. Furthermore, we also applied the automated ML toolkit ASC[11] as it has the potential to yield superior results owing to the automatic ensemble selection and hyperparameter optimization. For more information on this toolkit see Feurer et al. (2015). Finally, we evaluated the Riemannian-geometry-based decoding as implemented in the Python package pyRiemann[12] using an SVM with a RBF kernel, as it has recently achieved state-of-the art results in several BCI decoding tasks [Lotte et al. (2018)]. All models under investigation relied on implementations in scikit-learn[13] [Pedregosa et al. (2011)].

2.5. *Evaluation of performance*

We performed 5-fold cross-validation (CV) on the recordings of the development set, such that each recording was predicted exactly once. We did not shuffle data during splitting; rather, we used chronologically ordered splits. During CV, we optimized the hyperparameters of our feature-based models [see Table S1].

For final evaluation, we evaluated our models on the held back final evaluation set. We trained models on the full development set and predicted the examples in the final evaluation set. We repeated final evaluation five times to manage the statistical variances caused by initialization of certain models.

We report the accuracy score as $ACC = \frac{\#\ of\ correct\ predictions}{\#\ of\ examples}$ for the development and final evaluation sets as averages over the CV folds and final evaluation repetitions, respectively.

Furthermore, we compare model performance based on bootstrap accuracy distributions of final evaluation predictions. For each model we draw the same 10.000 bootstrap samples where each sample consists out of 100 randomly selected final evaluation predictions. We then compute the bootstrap accuracy and plot the resulting distribution.

In addition, we used a statistical sign test [Dixon and Mood (1946)] to validate the predicted labels in the final evaluation for superiority of model performance ($H1$). To provide a conservative estimate, occurring ties were equally split to both classes. We rejected the null hypothesis ($H0$: There is no difference in performance) at a p-value $<$0.05.

2.6. *Analysis*

Handcrafted features Because we implemented a large set of features [Table 3], we were interested in their individual importance to the decision process. RF estimates the importance of a feature by internally computing the "purity" of a data split obtained through a feature. In principal, the purer the data split and the earlier the feature is considered in the trees of the forest, the higher its importance. We assigned a textual label to all computed features described in Section 2.4 and mapped them to the average feature importance in CV using RF. We selected subsets of the features based on their textual labels with respect to frequency band and electrode location. We then created topological plots of the average feature importance in certain frequency ranges. Furthermore, we

[10] Available for download at https://github.com/alexandrebarachant/pyRiemann.

[11] Available for download at https://github.com/automl/auto-sklearn.

[12] See footnote 10.

[13] Available for download at https://scikit-learn.org/stable/.

computed the Spearman correlation of features over the development set and visualized the correlation map, as pronounced correlations could be a limiting factor to the interpretation of feature-based analyses.

Learned features We performed input-signal perturbation to determine the informative frequency ranges and electrode locations for identifying pathologies in the EEGs [Schirrmeister et al. (2017b)]. We computed the network predictions of the original and randomly perturbed input signals and correlated the amplitude change with the change of predictions. Given the labels of the examples, we could then determine whether an increase (or decrease) of signal amplitude in a given frequency range through perturbation contributed to more pathological (or non-pathological) predictions. Again, we made topological plots to indicate the most correlated frequency range and electrode location with the pathological class.

2.7. Ensembling

In ensembling, a number of model predictions are combined to provide an improvement over single-model performances. Under the assumption that models make uncorrelated errors, a combination can result in the overruling of incorrect single-model decisions. We computed the Spearman correlation of the CV predictions of all pairs of models and visualized the resulting correlation map. Furthermore, we computed and visualized the ratio of non-overlapping label errors of pairs of models to investigate the possibility of ensembling. Note that ASC was excluded from this investigation because it does not provide access to internally performed CV predictions. For ensembling, predictions are weighted. A weighting of “1” of individual model predictions is a special case and is equivalent to majority voting. We first built an ensemble of three models based on the highest ratio of non-overlapping label errors. We then computed an ensemble label based on the majority vote of labels computed from the individual model predictions. In addition, as an automated alternative, we investigated an ensemble selection technique based on auto-sklearn[14] introduced by Caruana et al. (2004) that automatically selects models for ensembling and computes the optimal weights based on the validation set. We evaluated the performance of both ensembles based on CV and final evaluation predictions of the individual models.

3. Results

3.1. Data descriptive statistics

We present histograms of the age distribution within the development and final evaluation sets of the TUH Abnormal EEG Corpus (v2.0.0) in Fig. 5. The age distribution, especially of the female patients, differed between the development and final evaluation sets. Moreover, the ratio of pathological and non-pathological examples also differed by gender and subset. The proportion of male and female patients, conversely, was closely matched. It can be observed that recordings labeled as pathological appear more frequent with higher age, which matches the intuition.

The observations are important in two respects. First, correlation of pathology with age could lead to a situation where the trained models use patient age as a proxy for pathology. To investigate the role of patient age in the decoding of EEG pathology further, we included age as a feature in a separate analysis (Section 3.6). Conversely, systematic differences between the development and final evaluation sets can reduce generalization and thus present a challenge. However, both correlation of pathology with age and shifts between the development and final evaluation data could occur in practical application scenarios. We hence considered these properties of the TUH data set as ecologically valid and methodologically interesting aspects. However, they must be considered when interpreting the results achieved on this data set.

3.2. Aggregated feature-based decoding

We present the aggregated feature-based decoding results in Fig. 6, right half. The Riemannian-geometry-based decoding achieved nearly 86% accuracy. Interestingly, the decoding accuracy increased from 81% in CV to 86% in the final evaluation, which could indicate underfitting of the training data in CV. We obtained greater than 84% accuracy using both a traditional and an automated feature-based approach.

Furthermore, we observed that all feature-based models had a higher ratio of false negatives than false positives; that is, they rather classified pathological examples as non-pathological than the opposite (Fig. 7, bottom row). This is in consensus with results presented by Lopez de Diego (2017), Schirrmeister et al. (2017a), and Van Leeuwen et al. (2019). Since the models show a lower sensitivity [Fig. 7], they miss to identify the EEG pathology of some patients. Especially for medical screening methods this is undesirable as people remain undiagnosed. Performance improvements would therefore be preferable [see Section 5].

To the best of our knowledge, there is only one other previously published feature-based result for binary pathology decoding based on the TUH Abnormal EEG corpus [Lopez de Diego (2017)]. It was achieved using cepstral coefficients and a CNN + MLP architecture for classification resulting in an accuracy of 78.8%. We thus increased this feature-based baseline by greater than 7% using RG, and more than 5% using RF and ASC.

Riemannian-geometry-based classification outperformed all other feature-based models and achieved 85.87% accuracy. We observed that treating covariance matrices with appropriate metrics of their native space (geometric instead of Euclidean mean) yielded superior performance [S4], which was to be expected [Barachant et al. (2013)].

This performance of the Riemannian-geometry-based decoding was remarkable considering that the covariance matrix of 21 electrodes had only 231 non-redundant entries. The covariance matrices, as well as our aggregated high-dimensional feature vectors, did not contain detailed time course information, and, in both cases, we averaged over the number of crops from which features were extracted. However, the results demonstrate that there is sufficient information contained in covariance matrices even to outperform all other tested models using handcrafted features. Similarly, for example Sabbagh et al. (2019) have also used Riemannian-geometry-based decoding and found strong results.

The application of ASC effectively reduces the required amount of time and expert knowledge required to build and optimize a well-working model and has the ultimate goal of making ML applicable by non-experts. We can confirm that, given our set of features, ASC achieved competitive results in classifying EEG pathology, without requiring user interaction. The ensemble that was automatically chosen by ASC consisted of AdaBoost [Schapire and Freund (2013)] with 78% (66%, 8% and 4%) of ensemble strength, gradient boosting with 18%, and linear discriminant analysis with 4%.

3.3. End-to-end decoding performance

We present the end-to-end decoding results with deep neural networks in Fig. 6, left half. Our overall best decoding result was 86.16% accuracy obtained by the BD-TCN. This accuracy was extremely close to the result of 86.57% accuracy previously reported using ChronoNet [Roy et al. (2019a)]. The BD-TCN was followed by BD-Deep4 and BD-Shallow with 84.57% and 84.13% accuracy, respectively. BD-EEGNet achieved a decoding accuracy of 83.41%. Interestingly, Heilmeyer et al. (2018) also found no statistically significant difference in accuracies comparing BD-Deep4 with BD-EEGNet in a large-scale benchmark test across different tasks and data sets. Overall, the networks did not demonstrate as much performance difference in the CV and final evaluation as the

[14] See footnote 11.

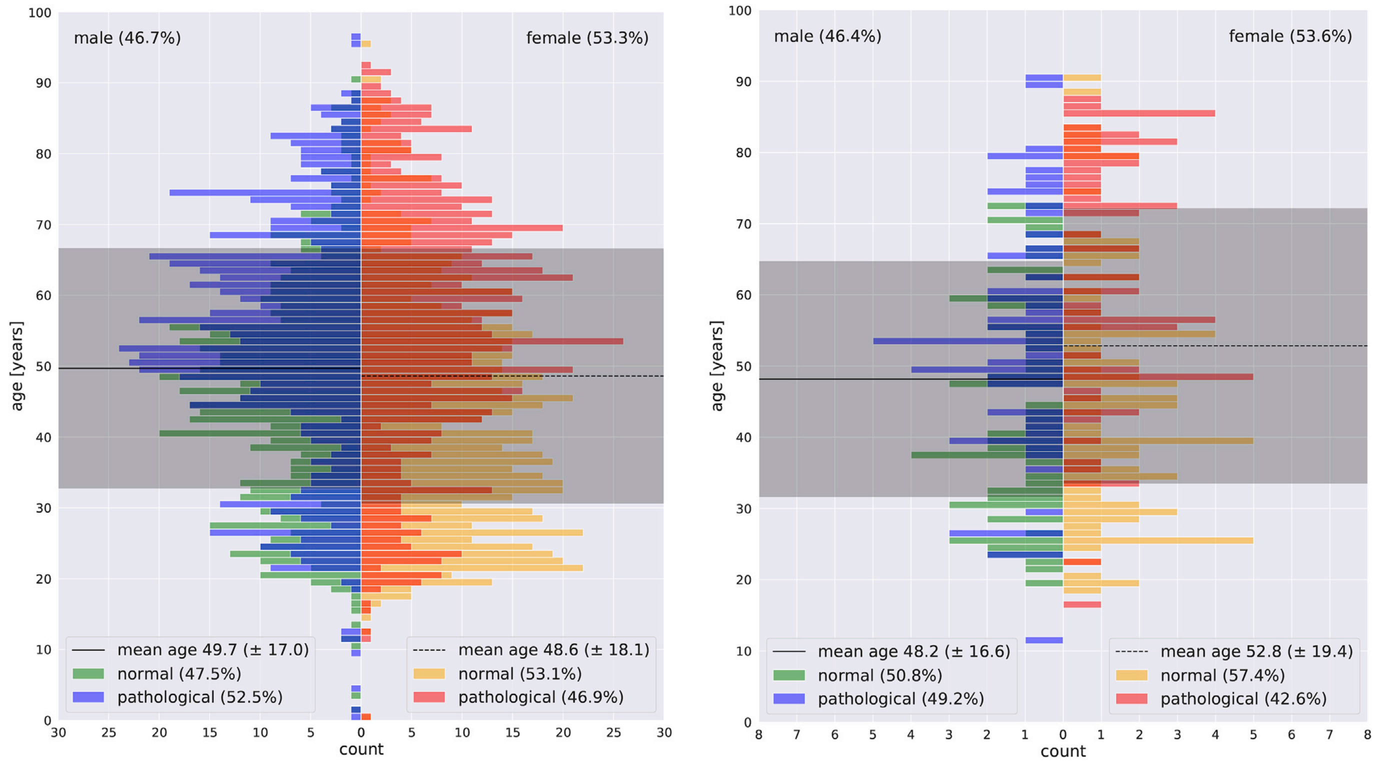


Fig. 5. Development (left) and final evaluation (right) subsets of TUH Abnormal EEG Corpus (v2.0.0). Histogram is constructed as an age pyramid subdivided into male and female patients. Different color coding indicates pathological and non-pathological EEG recordings.

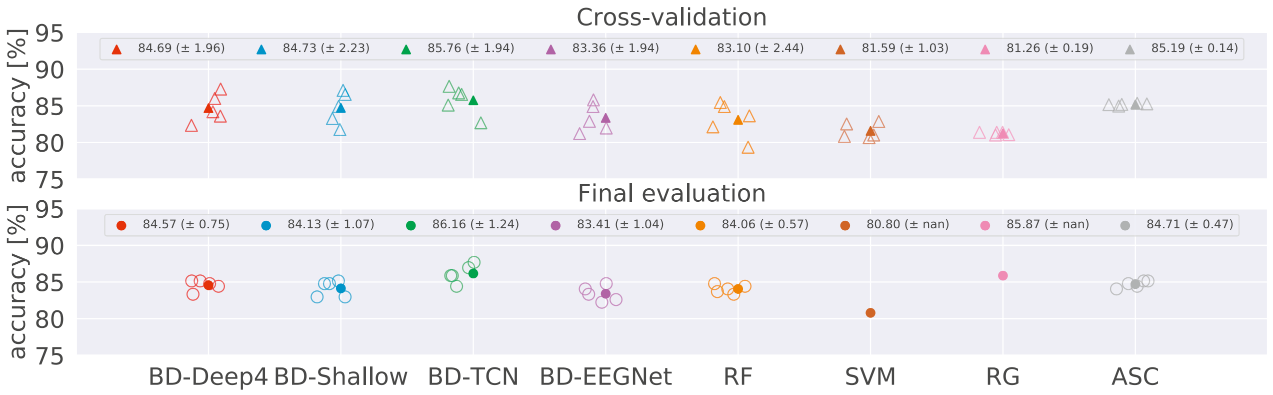


Fig. 6. Decoding accuracies of all models during CV and in final evaluation. TCN implemented in Braindecode (BD-TCN) indicated best performance. Decoding based on RG achieved accuracy similar to BD-TCN. BD-Deep4 and BD-Shallow ConvNets, RF, and ASC were on same level, whereas BD-EEGNet achieved marginally lower decoding accuracy. SVM indicated the worst performance.

feature-based approaches. For the networks, the differences were in the range of -0.6% to $+0.4\%$ and in the range of -1.21% to $+ 4,61\%$ for the feature-based approaches. Again, as in the feature-based approach, we observed more false negatives (Fig. 7, top row). Our observations thus suggest that the end-to-end methods used for our study may be more stable when comparing CV with final evaluation results.

We also tracked learning curves of all applied networks (Fig. 8). The loss and misclassification curves of BD-Deep4 and BD-Shallow are irregular at the start; however, they become smoother after Epoch 20. BD-EEGNet displayed smooth curves overall, however with the highest loss and misclassification rate. The curves of the BD-TCN indicate the greatest difference of training to test loss.

Our results indicate that the BD-TCN model is competitive with ChronoNet, which is a combined ConvNet/RNN architecture, for the given task in terms of decoding accuracy. The BD-TCN outperformed all other networks, which could be a consequence of its design and optimization through a neural architecture search [Chrabąszcz (2018)]. All other models were originally developed and optimized for other decoding tasks. Their performance in the present study underlines their general applicability.

The presented learning curves demonstrate astounding differences and we hypothesize that they are characteristic for the network architectures under investigation. The smoothing of the curves near Epoch 20, especially observable for BD-Deep4 and BD-Shallow, is likely the effect of cosine annealing updates of the learning rate. In all models, except BD-EEGNet, we can observe a clear difference between the training and test loss. We assume that BD-EEGNet is unable to better fit the training data owing to its relatively small number of parameters (see Table S3); its learning curves indeed indicate signs of underfitting. The opposite, overfitting, cannot be observed. This could be due to the regularization

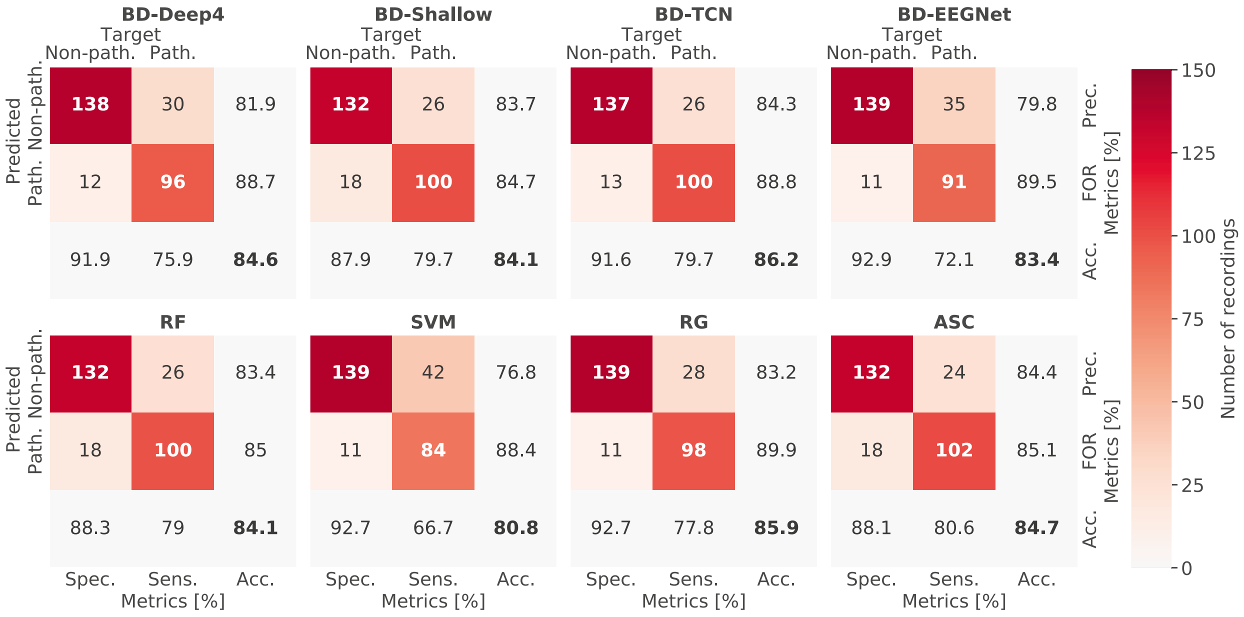


Fig. 7. Confusion matrices of all models averaged over independent final evaluation runs are presented in the upper left 2 × 2 sub-matrix. Specificity (Spec.), sensitivity (Sens.), precision (Prec.), false omission rate (FOR), and accuracy (Acc.) are indicated. All models determined more false negatives (pathological examples predicted as non-pathological).

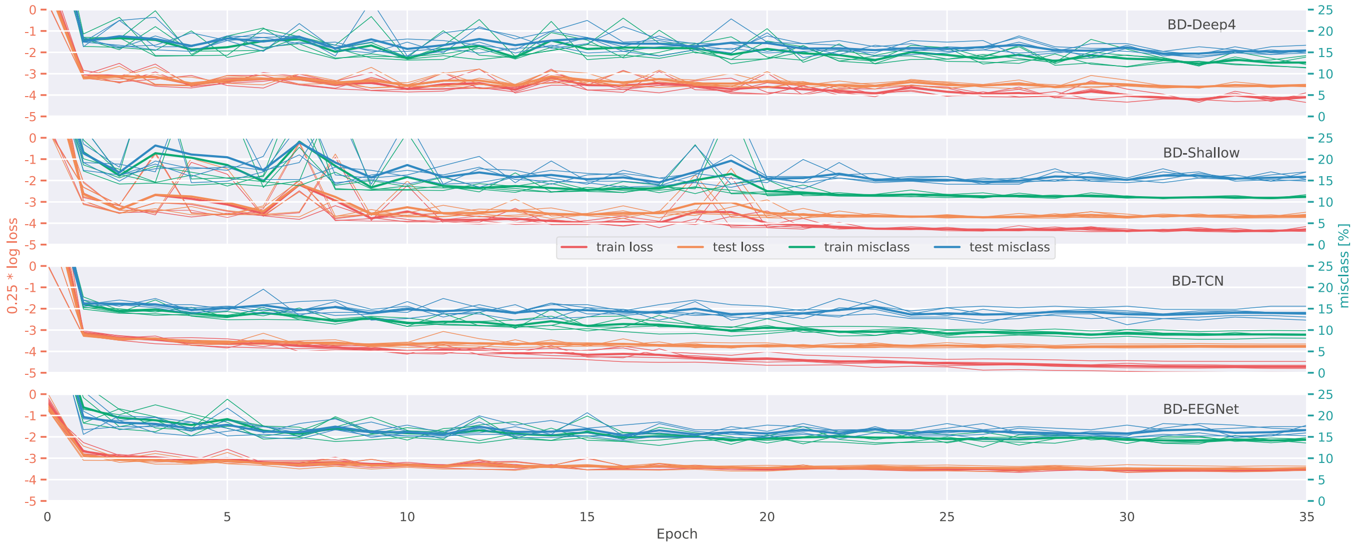


Fig. 8. Learning curves of investigated network architectures. Smoothing trend can be observed near Epoch 20, which could be effect of cosine annealing. BD-TCN achieves lowest misclassification rate. Curves of BD-EEGNet are a sign of underfitting.

techniques, i.e., dropout or weight decay.

3.4. Feature-based and end-to-end decoding performances in comparison

We observed decoding accuracies in the same range using different models and approaches, i.e., ChronoNet (86.57%), BD-TCN (86.16%), BD-Deep4 (84.57%), ASC (84.71%), RF (84.06%), and RG (85.87%). For further investigation of performance, we show the bootstrap accuracy distributions in Fig. 9. They are almost congruent, where the SVM is an exception. However, we did not find statistical evidence that one of the models under investigation performed better than the others (Fig. S1, lower triangle).

For a more in-depth analysis of model performance, we provide accuracies on subsets created based on age and gender information in Fig. 10. With minor exceptions, e.g. SVM on male patients younger than 30, models show very similar performance across all subsets. Overall, recordings of young female patients can be decoded with the highest accuracy, whereas old female patients seem to be the hardest subgroup. The middle aged group is the most consistent.

Previous publications [Table 1] have indicated that deep learning typically performs better in decoding pathology from EEGs using the TUH Abnormal EEG corpus. This is because there is only one published result using handcrafted features. It appears that this baseline was not particularly strong, which made deep learning approaches appear superior. However, herein, we determined similar decoding accuracies of feature-based and neural network approaches.

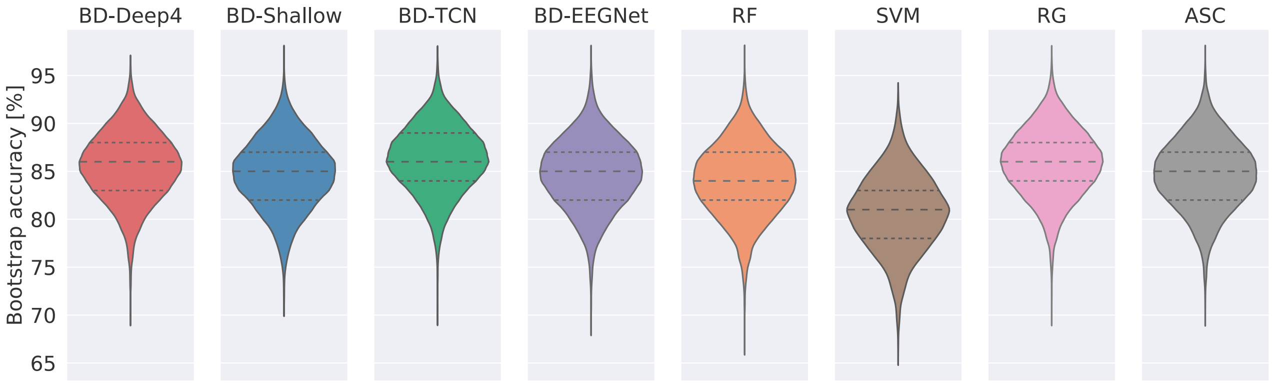


Fig. 9. Bootstrap accuracy distribution of all models under investigation. All distributions except for the SVM are almost congruent.

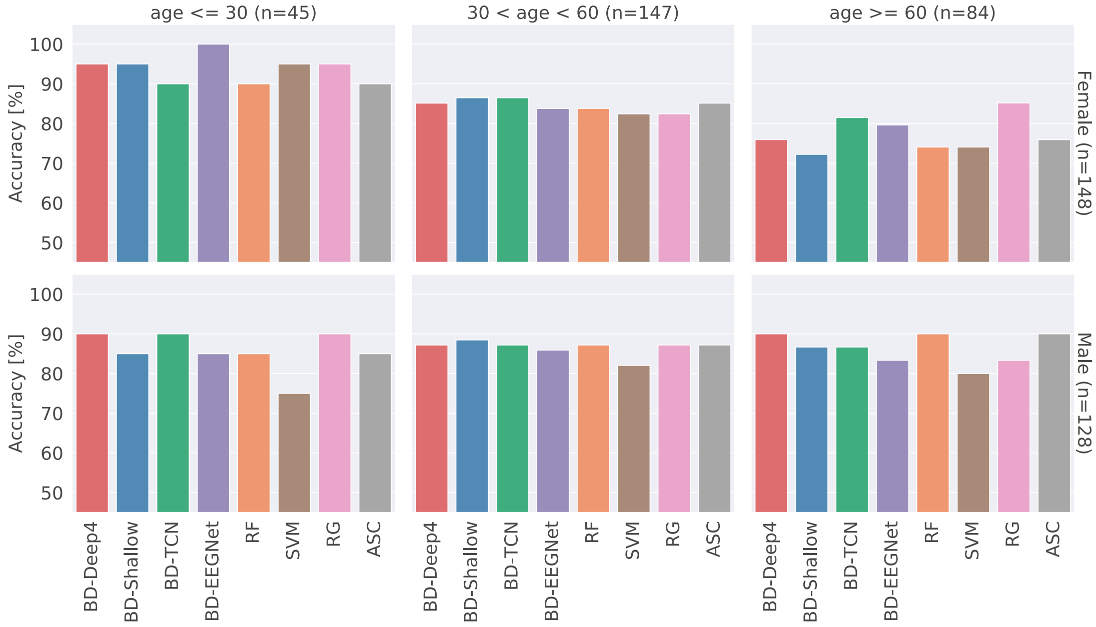


Fig. 10. Overview of average final evaluation accuracy performance in different subsets. Columns show different age categories, rows show the gender.

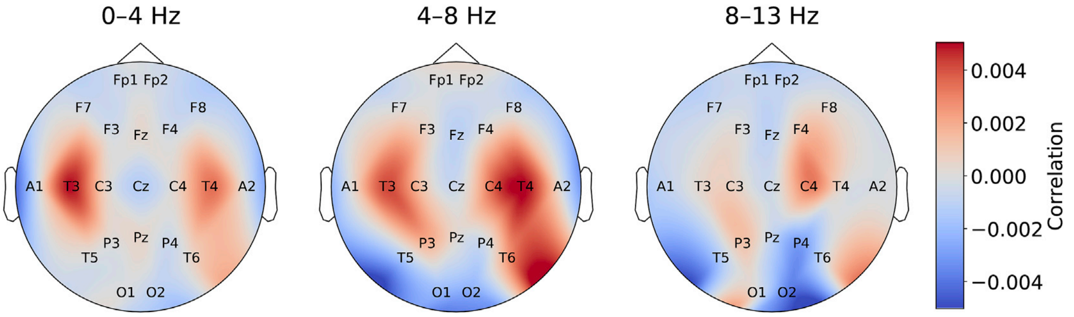


Fig. 11. Input-signal perturbation with BD-Deep4 network displays that higher activity at temporal electrode locations (T3, T4) is indicative of pathology, especially in low frequency ranges (0–4 Hz and 4–8 Hz). In alpha frequency range, there is negative correlation with pathological class at occipital electrode locations (O1, O2).

3.5. Importance of learned and handcrafted features

Through our perturbation analysis using the BD-Deep4, we determined correlations with predictions of the pathological class at temporal electrode locations T3 and T4 when increasing the amplitudes (Fig. 11). The effect was especially prominent in the delta and theta frequency range. Conversely, a decrease in correlation at occipital electrodes O1 and O2 was the most prominent effect in the alpha frequency range.

For a comparison with the perturbation analysis, we present the important handcrafted features extracted in the same frequency ranges using our analysis of feature importance using RF (Fig. 12). Features extracted at temporal electrode T4 in the delta and theta frequency range are most informative, which is in consensus with the perturbation result. Features extracted at electrode T3, however, are not considered as

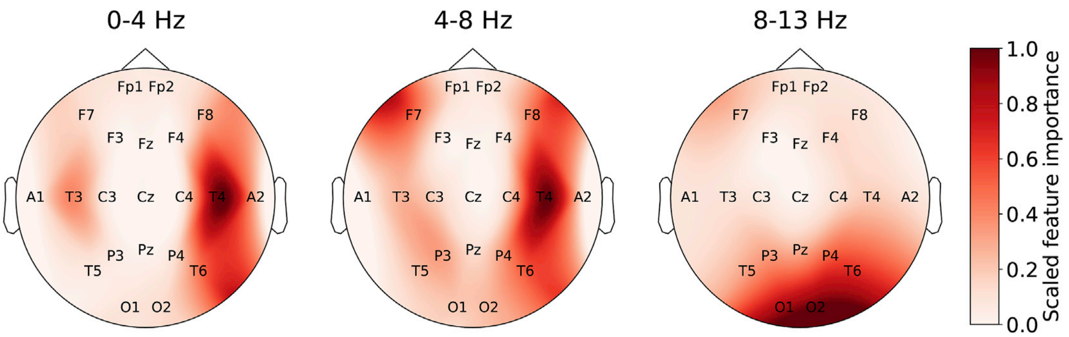


Fig. 12. RF feature importance analysis indicates that features extracted at 0–4 Hz and 4–8 Hz at temporal electrode T4 are most informative. In 8–13 Hz frequency band, occipital electrodes (O1, O2) have highest importance values.

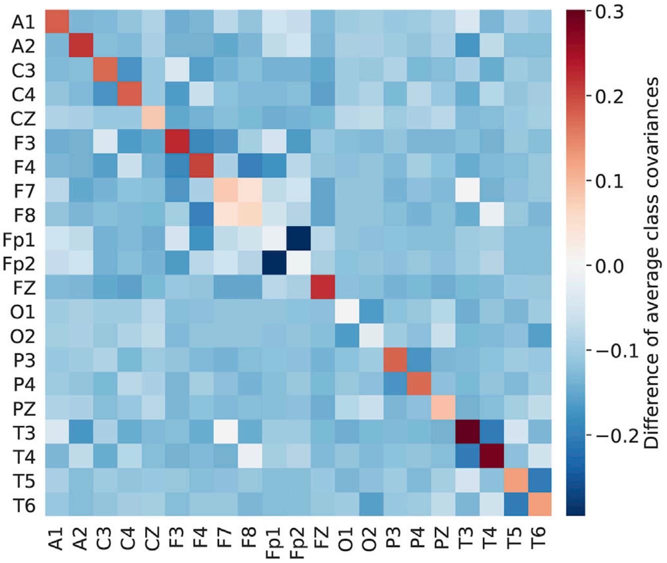


Fig. 13. Differences (pathological - non-pathological) of average class covariance matrices mapped to tangent space. Values most indicative of pathological activity are variance extracted at electrodes T3 and T4. Only variance of O1, O2, Fp1, and Fp2 is indicative of normal activity.

informative. In the alpha frequency range, the majority of the informative features were extracted at occipital electrodes O1 and O2. This, again, is in consensus with the perturbation result.

For a comparison with the perturbation and feature importance analysis, we present a visualization of the values in the matrix of differences of class covariance matrices that were used in the Riemannian-gemoetry-based decoding pipeline (Fig. 13). It can be observed that the variance extracted at temporal electrodes T3 and T4 is most indicative of a pathology. This is in consensus with the perturbation analysis. Furthermore, as in the perturbation analysis, variance at electrodes O1 and O2 (and Fp1 and Fp2) are indicative of normal brain activity. Both observations are underlined by the feature importance analysis.

To further analyze the different patterns of the feature importance and perturbation results, we computed correlations of features extracted from the delta, theta, and alpha range [Fig. 14]. The figure shows strong correlations, especially in the theta and alpha band. We actually observe strong correlations across all implemented features, despite feature domain, frequency band, and electrode recording site. [Fig. S2].

3.6. Time-resolved feature-based decoding

Inspired by the cropped-decoding setup used in our end-to-end pipeline, we implemented a time-resolved (non-aggregated) decoding for our feature-based models. However, this did not lead to an increase in decoding accuracy during CV, which is why we did not investigate this setup any further.

Given the drastic increase in the amount of data, it should be possible to realize an improvement using time-resolved decoding, if not here, then for a different task or data set. For the given data, the improvements could actually be negligible. Because the task is concerned with classifying an EEG recording as either pathological or non-pathological, there is likely no information in the signals that evolves at a large time scale. Because the TUH Abnormal EEG Corpus (v2.0.0) is not a seizure or event data set, we assume that if there exists a continuous alteration in brain activity reflecting a static dysfunction, e.g., related to a structural brain abnormality, it would be indicated consistently. This is also one of the key assumptions of our aggregated decoding approach, which performed as well as the ConvNets [see 3.1]. If this assumption did not hold, an

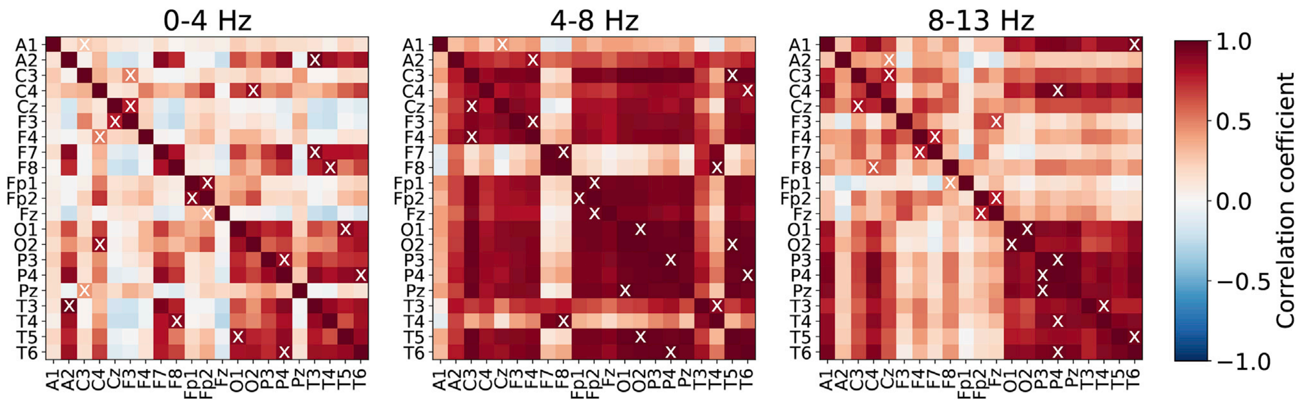


Fig. 14. Feature correlation analysis in frequency bands 0–4, 4–8, and 8–13 Hz. For every electrode (row), most highly correlated electrode is marked with a white cross.

aggregation, especially using the median as an aggregation function, would result in smoothing of the effect caused by the pathology. In a different data set, where signals change over a larger time scale, a time-resolved decoding could result in superior decoding performance. However, one must consider the challenges that are posed by time-resolved feature-based decoding. In our study, the amount of data increased by a factor of approximately 200 compared to aggregated decoding.

3.7. Effect of patient-specific information on decoding accuracies

We investigated the influence of age and gender of patients, with respect to the decoding accuracies, as these might be useful for classification. Therefore, we added the age and gender of the patients to our feature vectors and as additional network channels to our BD-TCN. Classification accuracy improved only marginally during CV, e.g., RF CV accuracy increased by 0.15% (from 83.1% to 83.25%) by adding age and gender. Interestingly, this is consistent with a recent publication, where the authors attempted to combine neural networks and the age information of patients to improve the decoding of pathology from EEGs [Van Leeuwen et al. (2019)]. They also found only marginal, insignificant improvements (+0.07 area under the receiver operating characteristic curve).

3.8. Ensemble decoding performance

We present the results of our two ensemble approaches in Table 4. The combination of the models BD-Deep4, RF, and RG had the highest ratio of non-overlapping CV label errors (336 errors, ratio of 44.56%), which is why we selected them for majority vote ensembling. See Table S5 for the number and ratio of non-overlapping errors of all three tuples of models. Automatic ensembling based on auto-sklearn chose every model except the SVM and computed optimal weights based on the CV predictions [Fig. 4].

Whereas the auto-sklearn ensemble resulted in the overall best CV accuracy (86.23%), neither ensembling based on majority voting (85.51%) nor automatic ensemble selection (85.14%) resulted in an improvement over best single-model performance (BD-TCN: 86.16%) in the final evaluation. Whereas the accuracy based on majority voting increased from CV to final evaluation, which can be an indication of underfitting, the accuracy based on auto-sklearn decreased, which can be an indication of overfitting.

4. Discussion

4.1. Deep end-to-end versus feature based-decoding accuracies

A main finding of our present study, together with the results of Van Leeuwen et al. (2019) and Roy et al. (2019a) is that EEG pathology decoding accuracies lie in a narrow range of 81–86%, even though we compared a broad range of:

- **Analysis strategies** including deep-end-to-end, feature-based, automated ML, and based on RG;
- **Network archetypes** including ConvNet, TCN, and RNN;
- **Network architectures** including BD-Deep4, BD-TCN, BD-EEGNet, BD-Shallow, and ChronoNet;
- **Feature-based classifiers and ensembles** including RF, SVM, and ASC, and;
- **Data sets** including the Temple University Hospital Abnormal EEG Corpus (v2.0.0) and the data set used by Van Leeuwen et al. (2019).

Importantly, this range was also considerably below a perfect classification score (100%). Decades of previous EEG research have indicated that inter-rater reliability in EEG diagnostics is only moderate [Grant et al. (2014), Houfek and Ellingson (1959); Rose et al. (1973)], which ultimately results in label noise. Regarding label noise, we refer to expert mistakes in diagnosing the EEG recordings as either pathological or non-pathological, which has a number of mutually related consequences [Frénay and Verleysen (2013)]: decrease of decoding performance, increase of required amount of data to achieve acceptable decoding performance, and increase of model complexity to properly fit the data. Furthermore, label noise complicates the identification of relevant features [Frénay and Verleysen (2013)]. The effect of label noise in EEG decoding is also discussed in Engemann et al. (2018) and Sun et al. (2019). Importantly, in our setting, low inter-rater agreement and the resulting label noise imposed a limit on the theoretically achievable decoding accuracies because we were required to evaluate against the noisy labels within our separate final evaluation set. Moreover, we did not have access to any rater-independent ground truth.[15] Interestingly, inter-rater agreement in binary classification of EEGs into pathological and non-pathological has been reported as 86–88% [Houfek and Ellingson (1959); Rose et al. (1973)], although these scores were based on EEG ratings of only two neurologists. Given these numbers, it would appear to be a possibility that EEG pathology decoding accuracies as observed herein and previously by Schirrmeister et al. (2017a), Roy et al. (2019a), and Van Leeuwen et al. (2019), at approximately 86% could have approached the theoretical optimum imposed by label noise. This proposed hypothesis could be tested in the future; however, it would involve a considerable effort. It would require both a large data set as used in the present study and independent ratings of multiple EEG experts. A massive amount of EEG data is waiting in the archives of medical centers to be used. More data would probably illicit positive effects, as more data in ML is typically favorable over more complex classification algorithms [Halevy et al. (2009)]. Furthermore, a labeling of EEG recordings as pathological or non-pathological through an ensemble of considerable number of neurologists and/or epileptologists would be a significant beginning to improve label quality. The inter-rater agreement scores that result from the ensemble could then be included in the data set on a per-recording basis, such that they can be included in a detailed analysis.

In the case where the hypothesis is incorrect and the theoretically optimal EEG pathology decoding accuracy is higher, we see two possible, nonexclusive causes. The first is that inter-rater reliability of the current data sets is higher than the numbers typically reported in the literature [see Houfek and Ellingson (1959); Rose et al. (1973)]. The second is that all the methods investigated to date did not extract or use certain features and information that was used by the physicians to determine a diagnosis. In the first case, the question arises why neither the end-to-end nor the feature-based pipeline could better fit and predict the data. In the second case, the question arises as to what additional source of information was used by the physicians and how it could be included to enhance performance. Both cases would open up new, interesting research questions.

4.2. Learned versus handcrafted features

Based on our feature visualizations, we determined that features extracted in the theta and delta range from the temporal electrode locations are considered informative. Knowing that epilepsy is statistically one of the most common disorders of the brain [Thijs et al. (2019)] and that temporal lobe epilepsy is the most prominent epilepsy [Helmstaedter and Elger (2009)], this could be a reason why this region is important in determining pathology in all decoding pipelines. Note that although we know epilepsy is one of the pathologies included in the data set, we neither know the exact number of occurrences nor how many of those suffer from temporal lobe epilepsy.

[15] One can design alternative decoding tasks different from pathology decoding based on the TUH Abnormal EEG Corpus, such as the decoding of patient gender or age.

Interestingly, only features extracted at electrode T4 are considered important in the theta band based on the feature importance analysis [Fig. 12]. This is in strong contrast to the network perturbation result [Fig. 11] and covariance matrix visualization [Fig. 13], where both hemispheres were considered equally informative in this range. Given the mechanisms of a decision tree, we assume that the tree chose features extracted at T4 in the theta band early in the decision process, because they are informative. Features extracted at T3 in the same frequency band were then not selected for further splitting of examples because they did not provide additional information. Hence, we assumed that the features were highly correlated. Our feature correlation analysis [Fig. 14] revealed several strong correlations of features extracted at different electrode locations and indeed, features extracted at T3 were most correlated with features extracted at T4 (correlation coefficient approximately 0.9) in frequency band 4–8 Hz .

Through the comparison of our analyses, we determined that the visualization of the feature importance [Fig. 12] was misleading, owing to strong feature correlations. Typically, ease of interpretation is considered an advantage of feature-based approaches over end-to-end approaches; however, we determined that there are also limitations to this interpretability. Pitfalls with RF feature correlations were also described by Altmann et al. (2010).

4.3. Clinical usefulness of current decoders of EEG pathology

Everyday medical applications typically require higher accuracy than the current state-of-the-art in EEG pathology decoding to be accepted. However, decoding pipelines with an accuracy in the current range can be valuable. For example, they could make EEG diagnostics available to patients that cannot attend specialized centers. This includes wide areas of developing countries where specialized centers and neurological experts are rare. Approximately 50 million people worldwide suffer from epilepsy [World Health Organization (2019)], of which the vast majority live in developing countries [World Health Organization (2019); Singh and Trevick (2016)]. In these countries, in addition to their disease or disorder, patients frequently suffer from social stigma [World Health Organization (2019); Newton and Garcia (2012)] owing to missing diagnoses and inexperience in addressing those diseases and disorders. In our opinion, although diagnostic accuracy is at approximately 86%, an automatic diagnosis is better than not being diagnosed at all. The pipelines could be used as a prescreening method, which could recommend a visit to a specialized center in the case of the detection of pathological activity.

4.4. Implications for EEG decoding evaluation in general

Our findings of statistically similar decoding accuracies of different networks and classifiers has implications for other publications. EEG decoding papers based on deep end-to-end learning frequently compare their results to only (rather) simple feature-based approaches or exclusively to other deep end-to-end learning results. For example, for the pathology decoding task based on the TUH Abnormal EEG corpus, all publications based on deep end-to-end learning compared to the result of Lopez de Diego (2017). This leads to the impression that deep neural networks heavily outperform feature-based approaches. However, in this study, we demonstrated that with a somewhat elaborate feature-based approach, one can achieve decoding results similar to deep end-to-end methods [Figs. 6, 7, 9 and 10]. To be more precise, there is actually no statistical evidence that the investigated networks perform any better than the feature-based approaches [Fig. S1]. In medical applications different from EEG decoding, such as functional magnetic resonance imaging (MRI) decoding, similar observations have been made [He et al. (2020); Schulz et al. (2019)]. For example, He et al. (2020) performed a comparison of traditional ML methods using kernel ridge regression and neural networks of different archetypes including ConvNets. They also found similar model performance when decoding numerous behavioral and demographic measures. We therefore emphasize once more, that fair comparisons to strong baselines are essential to assess the quality of decoding results. This does not only hold for EEG pathology decoding tasks, but also for EEG decoding tasks in general.

4.5. Public availability of resources and reproducibility

The availability of code that can be used to independently reproduce the published results is currently the only method to truly assess the scientific quality and generalizability of the proposed approaches. In their extensive review, Roy et al. (2019b) provide an overview of deep learning applications to EEG data. They list whether data and code bases of published studies are publicly available. Furthermore, they also discuss the importance of result reproduction for the entire research field. To the best of our knowledge, there are only six other published results for binary pathology decoding from EEGs, five of which are based on the TUH Abnormal EEG Corpus (Table 1). Van Leeuwen et al. (2019) have used and adapted our BD-Deep4 network and applied both the original and adapted version to the same task on a different, even larger data set. Their results indicate that both versions yield identical decoding results in a similar range (82% accuracy) to what we present in this study. Furthermore, they also investigated the effect of including patient age in the classification process and determined only marginal improvements, which is in consensus with our observations [see sec:result6]. To improve the reproducibility in EEG decoding further, we have uploaded the resources[16] of our previous study [Schirrmeister et al. (2017a)] and also uploaded the resources of our current study.[17]

4.6. Potential improvements of the decoding pipelines

Although we implemented a large set of features of different domains for this study, this collection is not close to completion. There are probably an infinite number of features one could implement, e.g., in the domain of connectivity features one could additionally investigate the usage of cross-correlation, cross-coherence, mutual information, omega complexity, s-estimator, and global field synchronization [Jalili et al. (2013)]. We have already attempted to add the gender and ages of the patients to improve classification [3.6]. In clinical diagnostics, physicians have access to even more information such as medical patient history and ongoing medication. It is an open challenge as to how to include this information in the decision process.

With extremely high-dimensional feature-spaces as presented in this study, it is natural to rethink dimensionality reduction methods. A small feature dimension is favorable because it yields shorter learning times and makes interpretation easier. Although our preliminary feature selection with PCA resulted in a decrease of decoding performance, there are several other approaches to be investigated, including independent component analysis [Comon (1994)] (ICA) or tensor decomposition [Sidiropoulos et al. (2017)]. Furthermore, one could also attempt to enhance signal quality by applying a source reconstruction method [Michel et al. (2004)] prior to the extraction of the features and classification. We have observed that features extracted from temporal electrode locations were highly informative for the decoding of EEG pathology [3.4]. A more precise localization of signals could also improve classification accuracy.

As in the implementation of features, there are numerous other feature-based classifiers available. ASC automatically selected models that we did not choose ourselves [Section 3.7], i.e., AdaBoost and Gradient Boosting. Both classifiers, however, are frequently implemented with decision trees that also form the basis of RFs. It could be worth to

[16] Available for download at https://github.com/robintibor/auto-eeg-diagnosis-example.

[17] Available for download at https://github.com/gemeinl/auto-eeg-diagnosis-comparison.

further investigate in the application of other classifiers. A specific case is when performing time-resolved feature-based decoding [Section 3.5], where the models chosen for the present study could not yet benefit from the drastic increase of data.

Finally, one could run a pipeline search optimization based on automated ML [Hutter et al. (2019)]. This could not only include feature scaling, selection of classifiers, and optimization of their hyperparameters as executed by ASC and as already performed for the present study. One could also attempt to optimize the hyperparameters for feature extraction itself, such as frequency bands, aggregation function, window length, window function, and others. We are convinced that a systematic optimization of these choices could lead to even better-performing feature-based EEG decoding results.

In the huge search space of network architectures, the models investigated in this study possibly lie in a subspace that contains well-performing architectures for this task. Most likely, these architectures are not yet optimal to decode EEG pathology. Many research groups are now promoting the development of neural architecture search (NAS) [Elsken et al. (2019)] to address the concern that handcrafting features is now being replaced by architecture crafting. NAS is achieving rapid progress and we assume that it will spawn even better-performing architectures in the close future.

5. Conclusion and outlook

The aim of this manuscript is not to imply that individual papers are inherently wrong, but to point towards misunderstandings that may arise in the research community from the combined results of these papers and to tackle these misunderstandings. Concretely, in prior literature on the TUH Abnormal EEG Corpus there had been no strong attempt to improve feature-based decoding. The only available result for feature-based methods was the initial result by Lopez de Diego (2017) and a linear method meant more as a sanity check [Schirrmeister et al. (2017a)]. In contrast, strong attempts were made to improve deep learning methods on this data set by [Schirrmeister et al. (2017a)], Roy et al. (2019a), Amin et al. (2019), and Alhussein et al. (2019). The overall effect of this can be the mentioned skewed picture of reality (DL strongly outperforming feature-based methods on this data set) and we attempt to correct this in this manuscript.

To extend the current approach beyond binary classification, we initiated work on restructuring the physician reports included in the TUH Abnormal EEG Corpus to a tabular format which is a more machine and user-friendly representation. The code for restructuring is uploaded to our repository.[18]

To circumvent the consequences of label noise in EEG pathology decoding, instead we propose to decode I) the gender of patients to better assess the potentials of feature-based and end-to-end pipelines, and II) the age of patients to use the gap of chronological age to predicted brain age as an alternative source for indication of pathology. In the literature this gap and its estimation based on MRI scans is referred to brainAGE [Franke (2013); Franke et al. (2010)].

Based on our present study, we see a promising future for automated EEG diagnostics. A well-working pipeline that implements the mentioned options for improvements could be helpful in the interpretation of EEG recordings. It could not only make EEG diagnostics available to patients that cannot attend specialized centers, but could also allow for earlier detection of pathologies on the same level as ensembles of human experts, and thereby somewhat reduce the global burden of neurological diseases and disorders.

Data and code availability statements

The TUH Abnormal EEG Corpus [Lopez de Diego (2017)] used for our study is a subset of the TUH EEG Corpus [Obeid and Picone (2016), https://doi.org/10.3389/fnins.2016.00196]. Both data sets are publicly available for download upon registration at www.isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml.

The code used for our study relies on the open source toolboxes Braindecode (github.com/TNTLFreiburg/braindecode) and Brainfeatures (github.com/TNTLFreiburg/brainfeatures). Code specific to the experiments performed for our study was uploaded to github.com/gemeinl/auto-eeg-diagnosis-comparison.

Declaration of competing interest

The authors declare no competing financial interests.

CRediT authorship contribution statement

Lukas A.W. Gemein: Software, Writing - original draft, Writing - review & editing, Visualization, Investigation. **Robin T. Schirrmeister:** Methodology, Software, Writing - review & editing, Validation. **Patryk Chrabąszcz:** Software, Writing - review & editing, Visualization. **Daniel Wilson:** Software, Writing - review & editing. **Joschka Boedecker:** Supervision. **Andreas Schulze-Bonhage:** Writing - review & editing. **Frank Hutter:** Conceptualization, Supervision, Writing - review & editing, Resources. **Tonio Ball:** Conceptualization, Writing - review & editing, Supervision, Project administration, Resources, Visualization.

Acknowledgements

This work was supported by the Freiburg Graduate School of Robotics, the German Ministry of Education and Research (BMBF) grant 13GW0053D (MOTOR-BIC), and the DFG grant EXC 1086 BrainLinks-BrainTools to the University of Freiburg, Germany.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.117021.

References

Albert, B., Zhang, J., Noyvirt, A., Setchi, R., Sjaaheim, H., Velikova, S., Strisland, F., 2016. Automatic EEG processing for the early diagnosis of traumatic brain injury. Procedia Compt. Sci. 96, 703–712.

Alhussein, M., Muhammad, G., Hossain, M.S., 2019. EEG pathology detection based on deep learning. IEEE Access 7, 27781–27788.

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26 (10), 1340–1347.

Amin, S.U., Hossain, M.S., Muhammad, G., Alhussein, M., Rahman, M.A., 2019. Cognitive smart healthcare for pathology detection and monitoring. IEEE Access 7, 10745–10753.

Ang, K.K., Chin, Z.Y., Zhang, H., Guan, C., 2008. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In: *2008* IEEE International Joint Conference on Neural Networks. IEEE World Congress on Computational Intelligence, pp. 2390–2397 (IEEE).

Bai, S., Kolter, J.Z., Koltun, V., 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling arXiv preprint arXiv:1803.01271.

Balli, T., Palaniappan, R., 2009. A combined linear & nonlinear approach for classification of epileptic EEG signals. In: 2009 4th International IEEE/EMBS Conference on Neural Engineering. IEEE, pp. 714–717.

Barachant, A., Bonnet, S., Congedo, M., Jutten, C., 2013. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. Neurocomputing 112, 172–178.

Biswal, S., Kulas, J., Sun, H., Goparaju, B., Westover, M.B., Bianchi, M.T., Sun, J., 2017. SLEEPNET: Automated Sleep Staging System via Deep Learning arXiv preprint arXiv: 1707.08262.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM, pp. 144–152.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Cai, H., Sha, X., Han, X., Wei, S., Hu, B., 2016. Pervasive EEG diagnosis of depression using deep belief network with three-electrodes EEG collector. In: *2016* IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 1239–1246.

[18] See footnote 17.

Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A., 2004. Ensemble selection from libraries of models. In: Proceedings of the Twenty-First International Conference on Machine Learning. ACM, p. 18.

Chrabąszcz, P., 2018. Neural Architecture Search. Master's thesis. Albert Ludwig University Freiburg.

Comon, P., 1994. Independent component analysis, a new concept? Signal Process. 36 (3), 287–314.

Craik, A., He, Y., Contreras-Vidal, J.L., 2019. Deep learning for electroencephalogram (EEG) classification tasks: a review. J. Neural. Eng. 16 (3), 031001.

Dixon, W.J., Mood, A.M., 1946. The statistical sign test. J. Am. Stat. Assoc. 41 (236), 557–566.

Elsken, T., Metzen, J.H., Hutter, F., 2019. Neural architecture search: a survey. J. Mach. Learn. Res. 20 (55), 1–21.

Engemann, D.A., Raimondo, F., King, J.-R., Rohaut, B., Louppe, G., Faugeras, F., Annen, J., Cassol, H., Gosseries, O., Fernandez-Slezak, D., et al., 2018. Robust EEG-based cross-site and cross-protocol classification of states of consciousness. Brain 141 (11), 3179–3192.

Esteller, R., Echauz, J., Tcheng, T., Litt, B., Pless, B., 2001. Line length: an efficient feature for seizure onset detection. In: *2001* Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 2. IEEE, pp. 1707–1710.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F., 2015. Efficient and robust automated machine learning. In: Advances in Neural Information Processing Systems, pp. 2962–2970.

Franke, K., 2013. BrainAge: a Novel Machine Learning Approach for Identifying Abnormal Age-Related Brain Changes. PhD thesis. University of Zurich.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., Initiative, A.D.N., et al., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage 50 (3), 883–892.

Frénay, B., Verleysen, M., 2013. Classification in the presence of label noise: a survey. IEEE Trans. Neural Netw. Learning Syst. 25 (5), 845–869.

Gemein, L.A.W., 2017. Automated EEG Diagnosis. Master's thesis. Albert Ludwig University Freiburg.

Giri, E.P., Fanany, M.I., Arymurthy, A.M., Wijaya, S.K., 2016. Ischemic stroke identification based on EEG and EOG using 1D convolutional neural network and batch normalization. In: *2016* International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, pp. 484–491.

Grant, A.C., Abdel-Baki, S.G., Weedon, J., Arnedo, V., Chari, G., Koziorynska, E., Lushbough, C., Maus, D., McSween, T., Mortati, K.A., et al., 2014. EEG interpretation reliability and interpreter confidence: a large single-center study. Epilepsy Behav. 32, 102–107.

Halevy, A., Norvig, P., Pereira, F., 2009. The unreasonable effectiveness of data. IEEE Intell. Syst. 24 (2), 8–12.

Hammer, J., Fischer, J., Ruescher, J., Schulze-Bonhage, A., Aertsen, A., Ball, T., 2013. The role of ECoG magnitude and phase in decoding position, velocity, and acceleration during continuous motor behavior. Front. Neurosci. 7, 200.

Hartmann, K.G., Schirrmeister, R.T., Ball, T., 2018. Hierarchical internal representation of spectral features in deep convolutional networks trained for EEG decoding. In: *2018* 6th International Conference on Brain-Computer Interface (BCI). IEEE, pp. 1–6.

He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T., 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. Neuroimage 206, 116276.

Heilmeyer, F.A., Schirrmeister, R.T., Fiederer, L.D., Volker, M., Behncke, J., Ball, T., 2018. A large-scale evaluation framework for EEG deep learning architectures. In: *2018* IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 1039–1045.

Helmstaedter, C., Elger, C.E., 2009. Chronic temporal lobe epilepsy: a neurodevelopmental or progressively dementing disease? Brain 132 (10), 2822–2830.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., et al., 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. 29.

Hjorth, B., 1970. EEG analysis based on time domain properties. Electroencephalogr. Clin. Neurophysiol. 29 (3), 306–310.

Hosseinifard, B., Moradi, M.H., Rostami, R., 2013. Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. Comput. Methods Progr. Biomed. 109 (3), 339–345.

Houfek, E.E., Ellingson, R.J., 1959. On the reliability of clinical EEG interpretation. J. Nerv. Ment. Dis. 128 (5), 425–437.

Hügle, M., Heller, S., Watter, M., Blum, M., Manzouri, F., Dumpelmann, M., Schulze-Bonhage, A., Woias, P., Boedecker, J., 2018. Early seizure detection with an energy-efficient convolutional neural network on an implantable microcontroller. In: 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–7.

Hutter, F., Kotthoff, L., Vanschoren, J. (Eds.), 2019. Automated Machine Learning: Methods, Systems, Challenges. Springer (in press), available at: http://automl.org/book.

Inouye, T., Shinosaki, K., Sakamoto, H., Toi, S., Ukai, S., Iyama, A., Katsuda, Y., Hirano, M., 1991. Quantification of EEG irregularity by use of the entropy of the power spectrum. Electroencephalogr. Clin. Neurophysiol. 79 (3), 204–210.

Jalili, M., Barzegaran, E., Knyazeva, M.G., 2013. Synchronization of EEG: bivariate and multivariate measures. IEEE Trans. Neural Syst. Rehabil. Eng. 22 (2), 212–221.

James, C.J., Lowe, D., 2003. Extracting multisource brain activity from a single electromagnetic channel. Artif. Intell. Med. 28 (1), 89–104.

Jasper, H., 1958. Report of the committee on methods of clinical examination in electroencephalography. Electroencephalogr. Clin. Neurophysiol. 10, 370–375.

Katz, M.J., 1988. Fractals and the analysis of waveforms. Comput. Biol. Med. 18 (3), 145–156.

Kemp, B., Värri, A., Rosa, A.C., Nielsen, K.D., Gade, J., 1992. A simple format for exchange of digitized polygraphic recordings. Electroencephalogr. Clin. Neurophysiol. 82 (5), 391–393.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization arXiv preprint arXiv:1412.6980.

Kiral-Kornek, I., Roy, S., Nurse, E., Mashford, B., Karoly, P., Carroll, T., Payne, D., Saha, S., Baldassano, S., O'Brien, T., Grayden, D., Cook, M., Freestone, D., Harrer, S., 2018. Epileptic seizure prediction using big data and deep learning: toward a mobile system. EBioMedicine 27, 103–111.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Kuhlmann, L., Cook, M., Fuller, K., Grayden, D., Burkitt, A., Mareels, I., 2008. Correlation analysis of seizure detection features. *2008* International Conference on Intelligent Sensors, Sensor Networks and Information Processing. IEEE, pp. 309–314.

Kumar, S.P., Sriraam, N., Benakop, P., Jinaga, B., 2010. Entropies based detection of epileptic seizures with artificial neural network classifiers. Expert Syst. Appl. 37 (4), 3284–3291.

Lachaux, J.-P., Rodriguez, E., Martinerie, J., Varela, F.J., 1999. Measuring phase synchrony in brain signals. Hum. Brain Mapp. 8 (4), 194–208.

Landis, J.R., Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. biometrics, pp. 159–174.

Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J., 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. J. Neural. Eng. 15 (5), 056013.

LeCun, Y., Haffner, P., Bottou, L., Bengio, Y., 1999. Object recognition with gradient-based learning. In: Shape, Contour and Grouping in Computer Vision. Springer, pp. 319–345.

Lehmann, C., Koenig, T., Jelic, V., Prichep, L., John, R.E., Wahlund, L.-O., Dodge, Y., Dierks, T., 2007. Application and comparison of classification algorithms for recognition of alzheimer's disease in electrical brain activity (eeg). J. Neurosci. Methods 161 (2), 342–350.

Logesparan, L., Casson, A.J., Rodriguez-Villegas, E., 2012. Optimal features for online seizure detection. Med. Biol. Eng. Comput. 50 (7), 659–669.

Lopez de Diego, S., 2017. Automated Interpretation of Abnormal Adult Electroencephalography. Master's thesis. Temple University.

Loshchilov, I., Hutter, F., 2016. SGDR: Stochastic Gradient Descent with Warm Restarts arXiv preprint arXiv:1608.03983.

Loshchilov, I., Hutter, F., 2017. Fixing Weight Decay Regularization in Adam arXiv preprint arXiv:1711.05101.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., Yger, F., 2018. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. J. Neural. Eng. 15 (3), 031005.

Michel, C.M., Murray, M.M., Lantz, G., Gonzalez, S., Spinelli, L., de Peralta, R.G., 2004. EEG source imaging. Clin. Neurophysiol. 115 (10), 2195–2222.

Minasyan, G.R., Chatten, J.B., Chatten, M.J., Harner, R.N., 2010. Patient-specific early seizure detection from scalp EEG. J. Clin. Neurophysiol.: Off. Pub. Am. Electroencephal. Soc. 27 (3), 163.

Mirowski, P., Madhavan, D., LeCun, Y., Kuzniecky, R., 2009. Classification of patterns of EEG synchronization for seizure prediction. Clin. Neurophysiol. 120 (11), 1927–1940.

Montavon, G., Samek, W., Müller, K.-R., 2018. Methods for interpreting and understanding deep neural networks. Digit. Signal Process. 73, 1–15.

Müller-Gerking, J., Pfurtscheller, G., Flyvbjerg, H., 1999. Designing optimal spatial filters for single-trial EEG classification in a movement task. Clin. Neurophysiol. 110 (5), 787–798.

Newton, C.R., Garcia, H.H., 2012. Epilepsy in poor regions of the world. Lancet 380 (9848), 1193–1201.

Obeid, I., Picone, J., 2016. The Temple university hospital EEG data corpus. Front. Neurosci. 10.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12 (Oct), 2825–2830.

Peng, C.-K., Havlin, S., Stanley, H.E., Goldberger, A.L., 1995. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos: Interdis. J. Nonlinear Sci. 5 (1), 82–87.

Petrosian, A., 1995. Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns. In: Proceedings Eighth IEEE Symposium on Computer-Based Medical Systems. IEEE, pp. 212–217.

Picone, J., 2019. Comment on Inter-rater Agreement in the TUH Abnormal EEG Corpus. Personal communication.

Quiroga, R.Q., Blanco, S., Rosso, O., Garcia, H., Rabinowicz, A., 1997. Searching for hidden information with Gabor Transform in generalized tonic-clonic seizures. Electroencephalogr. Clin. Neurophysiol. 103 (4), 434–439.

Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., et al., 2018. Scalable and accurate deep learning with electronic health records. NPJ Digital Med. 1 (1), 18.

Roberts, S.J., Penny, W., Rezek, I., 1999. Temporal and spatial complexity measures for electroencephalogram based brain-computer interfacing. Med. Biol. Eng. Comput. 37 (1), 93–98.

Rose, S.W., Penry, J.K., White, B.G., Sato, S., 1973. Reliability and validity of visual EEG assessment in third grade children. Clin. Electroencephalogr. 4 (4), 197–205.

Roy, S., Kiral-Kornek, I., Harrer, S., 2019a. ChronoNet: a deep recurrent neural network for abnormal EEG identification. In: Conference on Artificial Intelligence in Medicine in Europe. Springer, pp. 47–56.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J., 2019b. Deep learning-based electroencephalography analysis: a systematic review. J. Neural. Eng. 16 (5), 051001.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., et al., 1988. Learning representations by back-propagating errors. Cognit. Model. 5 (3), 1.

Sabbagh, D., Ablin, P., Varoquaux, G., Gramfort, A., Engemann, D.A., 2019. Manifold-regression to predict from MEG/EEG brain signals without source modeling. In: Advances in Neural Information Processing Systems, pp. 7321–7332.

Schapire, R.E., Freund, Y., 2013. Boosting: Foundations and Algorithms. Kybernetes.

Schirrmeister, R.T., Gemein, L., Eggensperger, K., Hutter, F., Ball, T., 2017a. Deep Learning with Convolutional Neural Networks for Decoding and Visualization of EEG Pathology arXiv preprint arXiv:1708.08012.

Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T., 2017b. Deep learning with convolutional neural networks for EEG decoding and visualization. Hum. Brain Mapp. 38 (11), 5391–5420.

Schulz, M.-A., Yeo, T., Vogelstein, J., Mourao-Miranada, J., Kather, J., Kording, K., Richards, B.A., Bzdok, D., 2019. Deep Learning for Brains?: Different Linear and Nonlinear Scaling in UK Biobank Brain Images vs. Machine-Learning Datasets. bioRxiv, p. 757054.

Sidiropoulos, N.D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E.E., Faloutsos, C., 2017. Tensor decomposition for signal processing and machine learning. IEEE Trans. Signal Process. 65 (13), 3551–3582.

Singh, A., Trevick, S., 2016. The epidemiology of global epilepsy. Neurol. Clin. 34 (4), 837–847.

Sturm, I., Lapuschkin, S., Samek, W., Müller, K.-R., 2016. Interpretable deep neural networks for single-trial EEG classification. J. Neurosci. Methods 274, 141–145.

Subasi, A., 2007. EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Syst. Appl. 32 (4), 1084–1093.

Subasi, A., Kevric, J., Abdullah Canbaz, M., 2019. Epileptic seizure detection using hybrid machine learning methods. Neural Comput. Appl. 31 (1), 317–325.

Sun, H., Kimchi, E., Akeju, O., Nagaraj, S.B., McClain, L.M., Zhou, D.W., Boyle, E., Zheng, W.-L., Ge, W., Westover, M.B., 2019. Automated tracking of level of consciousness and delirium in critical illness using deep learning. NPJ Digital Med. 2 (1), 1–8.

Thijs, R.D., Surges, R., O'Brien, T.J., Sander, J.W., 2019. Epilepsy in adults. Lancet 393 (10172), 689–701.

Van Leeuwen, K., Sun, H., Tabaeizadeh, M., Struck, A., Van Putten, M., Westover, M., 2019. Detecting abnormal electroencephalograms using deep convolutional networks. Clin. Neurophysiol. 130 (1), 77–84.

van Putten, M.J., Kind, T., Visser, F., Lagerburg, V., 2005. Detecting temporal lobe seizures from scalp EEG recordings: a comparison of various features. Clin. Neurophysiol. 116 (10), 2480–2489.

Watter, M., 2014. Epileptic Seizure Detection with Reservoir Computing. Master's thesis. Albert Ludwig University Freiburg.

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometr. Intell. Lab. Syst. 2 (1–3), 37–52.

World Health Organization, W., 2019. Epilepsy: a Public Health Imperative: Summary. World Health Organization. Technical report.