Deep learning-based electroencephalography analysis: a systematic review

OPEN ACCESS
IOP Publishing
Journal of Neural Engineering
J. Neural Eng. **16** (2019) 051001 (37pp)
https://doi.org/10.1088/1741-2552/ab260c

Topical Review

Deep learning-based electroencephalography analysis: a systematic review

Yannick Roy[1,5], **Hubert Banville**[2,3,5], **Isabela Albuquerque**[4], **Alexandre Gramfort**[2], **Tiago H Falk**[4] **and Jocelyn Faubert**[1]

[1] Faubert Lab, Université de Montréal, Montréal, Canada
[2] Inria, Université Paris-Saclay, Paris, France
[3] InteraXon Inc., Toronto, Canada
[4] MuSAE Lab, INRS-EMT, Université du Québec, Montréal, Canada

E-mail: yannick.roy@umontreal.ca

Received 20 February 2019, revised 30 May 2019
Accepted for publication 31 May 2019
Published 14 August 2019

Abstract

Context. Electroencephalography (EEG) is a complex signal and can require several years of training, as well as advanced signal processing and feature extraction methodologies to be correctly interpreted. Recently, deep learning (DL) has shown great promise in helping make sense of EEG signals due to its capacity to learn good feature representations from raw data. Whether DL truly presents advantages as compared to more traditional EEG processing approaches, however, remains an open question. *Objective.* In this work, we review 154 papers that apply DL to EEG, published between January 2010 and July 2018, and spanning different application domains such as epilepsy, sleep, brain–computer interfacing, and cognitive and affective monitoring. We extract trends and highlight interesting approaches from this large body of literature in order to inform future research and formulate recommendations. *Methods.* Major databases spanning the fields of science and engineering were queried to identify relevant studies published in scientific journals, conferences, and electronic preprint repositories. Various data items were extracted for each study pertaining to (1) the data, (2) the preprocessing methodology, (3) the DL design choices, (4) the results, and (5) the reproducibility of the experiments. These items were then analyzed one by one to uncover trends. *Results.* Our analysis reveals that the amount of EEG data used across studies varies from less than ten minutes to thousands of hours, while the number of samples seen during training by a network varies from a few dozens to several millions, depending on how epochs are extracted. Interestingly, we saw that more than half the studies used publicly available data and that there has also been a clear shift from intra-subject to inter-subject approaches over the last few years. About 40% of the studies used convolutional neural networks (CNNs), while 13% used recurrent neural networks (RNNs), most often with a total of 3–10 layers. Moreover, almost one-half of the studies trained their models on raw or preprocessed EEG time series.

[5] The first two authors contributed equally to this work.

Finally, the median gain in accuracy of DL approaches over traditional baselines was 5.4% across all relevant studies. More importantly, however, we noticed studies often suffer from poor reproducibility: a majority of papers would be hard or impossible to reproduce given the unavailability of their data and code. *Significance.* To help the community progress and share work more effectively, we provide a list of recommendations for future studies and emphasize the need for more reproducible research. We also make our summary table of DL and EEG papers available and invite authors of published work to contribute to it directly. A planned follow-up to this work will be an online public benchmarking portal listing reproducible results.

Keywords: EEG, electroencephalogram, deep learning, review, neural networks, survey

(Some figures may appear in colour only in the online journal)

1. Introduction

1.1. Measuring brain activity with EEG

Electroencephalography (EEG), the measure of the electrical fields produced by the active brain, is a brain mapping and neuroimaging technique widely used inside and outside the clinical domain [22, 74, 169]. Specifically, EEG picks up the electric potential differences, on the order of tens of μV, that reach the scalp when tiny excitatory post-synaptic potentials produced by pyramidal neurons in the cortical layers of the brain sum together. The potentials measured therefore reflect neuronal activity and can be used to study a wide array of brain processes.

Thanks to the great speed at which electric fields propagate, EEG has an excellent temporal resolution: events occurring at millisecond timescales can typically be captured. However, EEG suffers from low spatial resolution, as the electric fields generated by the brain are smeared by the tissues, such as the skull, situated between the sources and the sensors. As a result, EEG channels are often highly correlated spatially. The source localization problem, or inverse problem, is an active area of research in which algorithms are developed to reconstruct brain sources given EEG recordings [78].

There are many applications for EEG. For example, in clinical settings, EEG is often used to study sleep patterns [1] or epilepsy [3]. Various conditions have also been linked to changes in electrical brain activity, and can therefore be monitored to various extents using EEG. These include attention deficit hyperactivity disorder (ADHD) [11], disorders of consciousness [48, 54], depth of anaesthesia [68], etc. EEG is also widely used in neuroscience and psychology research, as it is an excellent tool for studying the brain and its functioning. Applications such as cognitive and affective monitoring are very promising as they could allow unbiased measures of, for example, an individual's level of fatigue, mental workload, [21, 195], mood, or emotions [5]. Finally, EEG is widely used in brain–computer interfaces (BCIs)—communication channels that bypass the natural output pathways of the brain—to allow brain activity to be directly translated into directives that affect the user's environment [117].

1.2. Current challenges in EEG processing

Although EEG has proven to be a critical tool in many domains, it still suffers from a few limitations that hinder its effective analysis or processing. First, EEG has a low signal-to-noise ratio (SNR) [23, 86], as the brain activity measured is often buried under multiple sources of environmental, physiological and activity-specific noise of similar or greater amplitude called 'artifacts'. Various filtering and noise reduction techniques have to be used therefore to minimize the impact of these noise sources and extract true brain activity from the recorded signals.

EEG is also a non-stationary signal [37, 65], that is its statistics vary across time. As a result, a classifier trained on a temporally-limited amount of user data might generalize poorly to data recorded at a different time on the same individual. This is an important challenge for real-life applications of EEG, which often need to work with limited amounts of data.

Finally, high inter-subject variability also limits the usefulness of EEG applications. This phenomenon arises due to physiological differences between individuals, which vary in magnitude but can severely affect the performance of models that are meant to generalize across subjects [36]. Since the ability to generalize from a first set of individuals to a second, unseen set is key to many practical applications of EEG, a lot of effort is being put into developing methods that can handle inter-subject variability.

To solve some of the above-mentioned problems, processing pipelines with domain-specific approaches are often used. A significant amount of research has been put into developing processing pipelines to clean, extract relevant features, and classify EEG data. State-of-the-art techniques, such as Riemannian geometry-based classifiers and adaptive classifiers [116], can handle these problems with varying levels of success.

Additionally, a wide variety of tasks would benefit from a higher level of automated processing. For example, sleep scoring, the process of annotating sleep recordings by categorizing windows of a few seconds into sleep stages, currently requires a lot of time, being done manually by trained technicians. More sophisticated automated EEG processing could make this process much faster and more flexible. Similarly, real-time detection or prediction of the onset of an epileptic

seizure would be very beneficial to epileptic individuals, but also requires automated EEG processing. For each of these applications, most common implementations require domain-specific processing pipelines, which further reduces the flexibility and generalization capability of current EEG-based technologies.

1.3. Improving EEG processing with deep learning

To overcome the challenges described above, new approaches are required to improve the processing of EEG towards better generalization capabilities and more flexible applications. In this context, deep learning (DL) [98] could significantly simplify prcessing pipelines by allowing automatic end-to-end learning of preprocessing, feature extraction and classification modules, while also reaching competitive performance on the target task. Indeed, in the last few years, DL architectures have been very successful in processing complex data such as images, text and audio signals [98], leading to state-of-the-art performance on multiple public benchmarks—such as the large scale visual recognition challenge [42]—and an ever-increasing role in industrial applications.

DL, a subfield of machine learning, studies computational models that learn hierarchical representations of input data through successive non-linear transformations [98]. Deep neural networks (DNNs), inspired by earlier models such as the perceptron [155], are models where: (1) stacked layers of artificial 'neurons' each apply a linear transformation to the data they receive and (2) the result of each layer's linear transformation is fed through a non-linear activation function. Importantly, the parameters of these transformations are learned by directly minimizing a cost function. Although the term 'deep' implies the inclusion of many layers, there is no consensus on how to measure depth in a neural network and therefore on what really constitutes a deep network and what does not [61].

Figure 1 presents an overview of how EEG data (and similar multivariate time series) can be formatted to be fed into a DL model, along with some important terminology (see section 1.4), as well as an illustration of a generic neural network architecture. Usually, when c channels are available and a window has length l samples, the input of a neural network for EEG processing consists of an array $X_i \in \mathbb{R}^{c\times l}$ containing the l samples corresponding to a window for all channels. This two-dimensional array can be used directly as an example for training a neural network, or could first be unrolled into a n-dimensional array (where $n = c \times l$) as shown in figure 1(b). As for the m-dimensional output, it could represent the number of classes in a multi-class classification problem. Variations of this end-to-end formulation can be imagined where the window X_i is first passed through a preprocessing and feature extraction pipeline (e.g. time-frequency transform), yielding an example X_i' which is then used as input to the neural network instead.

Different types of layers are used as building blocks in neural networks. Most commonly, those are fully-connected (FC), convolutional or recurrent layers. We refer to models using these types of layers as FC networks, convolutional neural networks (CNNs) [99] and recurrent neural networks (RNNs) [159], respectively. Here, we provide a quick overview of the main architectures and types of models. The interested reader is referred to the relevant literature for more in-depth descriptions of DL methodology [61, 98, 168].

FC layers are composed of fully-connected neurons, i.e. where each neuron receives as input the activations of every single neuron of the preceding layer. Convolutional layers, on the other hand, impose a particular structure where neurons in a given layer only see a subset of the activations of the preceding one. This structure, akin to convolutions in signal or image processing from which it gets its name, encourages the model to learn invariant representations of the data. This property stems from another fundamental characteristic of convolutional layers, which is that parameters are shared across different neurons—this can be interpreted as if there were filters looking for the same information across patches of the input. In addition, pooling layers can be introduced, such that the representations learned by the model become invariant to slight translations of the input. This is often a desirable property: for instance, in an object recognition task, translating the content of an image should not affect the prediction of the model. Imposing these kinds of priors thus works exceptionally well on data with spatial structure. In contrast to convolutional layers, recurrent layers impose a structure by which, in its most basic form, a layer receives as input both the preceding layer's current activations and its own activations from a previous time step. Models composed of recurrent layers are thus encouraged to make use of the temporal structure of data and have shown high performance in natural language processing (NLP) tasks [229, 249].

Additionally, outside of purely supervised tasks, other architectures and learning strategies can be built to train models when no labels are available. For example, autoencoders (AEs) learn a representation of the input data by trying to reproduce their input given some constraints, such as sparsity or the introduction of artificial noise [61]. Generative adversarial networks (GANs) [62] are trained by opposing a generator (G), that tries to generate fake examples from an unknown distribution of interest, to a discriminator (D), that tries to identify whether the input it receives has been artificially generated by G or is an example from the unknown distribution of interest. This dynamic can be compared to the one between a thief (G) making fake money and the police (D) trying to distinguish fake money from real money. Both agents push one another to get better, up to a point where the fake money looks exactly like real money. The training of G and D can thus be interpreted as a two-player zero-sum minimax game. When equilibrium is reached, the probability distribution approximated by G converges to the real data distribution [62].

Overall, there are multiple ways in which DL improve and extend existing EEG processing methods. First, the hierarchical nature of DNNs means features could potentially be learned on raw or minimally preprocessed data, reducing the need for domain-specific processing and feature extraction pipelines. Features learned through a DNN might also be more effective or expressive than the ones engineered by humans. Second, as is the case in the multiple domains

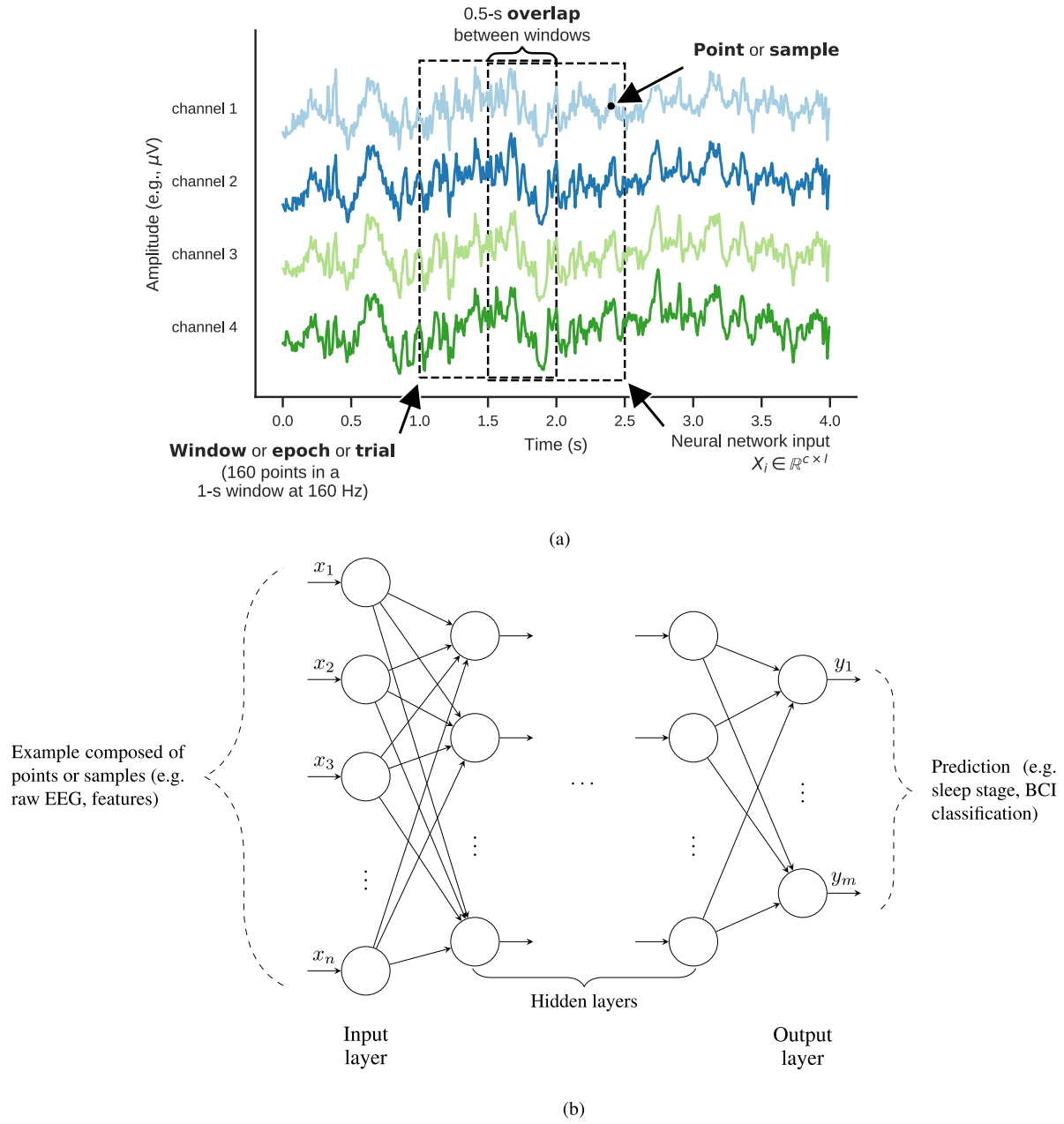

Figure 1. Deep learning-based EEG processing pipeline and related terminology. (a) Overlapping windows (which may correspond to trials or epochs in some cases) are extracted from multichannel EEG recordings. (b) Illustration of a general neural network architecture.

where DL has surpassed the previous state-of-the-art, it has the potential to produce higher levels of performance on different analysis tasks. Third, DL facilitates the development of tasks that are less often attempted on EEG data such as generative modelling [60] and domain adaptation [18]. The use of deep learning-based methods allowed the synthesis of high-dimensional structured data such as images [28] and speech [136]. Generative models can be leveraged to learn intermediate representations or for data augmentation [60]. In the case of domain adaptation, the use deep neural networks along with techniques such as correlation alignment [184] allows the end-to-end learning of domain-invariant representations, while preserving task-dependent information. Similar strategies can also be applied to EEG data in order to learn better representations and thus improve the performance of EEG-based models across different subjects and tasks.

On the other hand, there are various reasons why DL might not be optimal for EEG processing and that may justify the skepticism of some of the EEG community. First and foremost, the datasets typically available in EEG research contain far fewer examples than what has led to the current state-of-the-art in DL-heavy domains such as computer vision (CV) and NLP. Data collection being relatively expensive and data accessibility often being hindered by privacy concerns—especially with clinical data—openly available datasets of similar sizes are not common. Some initiatives have tried to tackle this problem though [73]. Second, the peculiarities of EEG, such as its low SNR, make EEG data different from

Table 1. Disambiguation of common terms used in this review.

	Definition used in this review
Point or sample	A measure of the instantaneous electric potential picked up by the EEG sensors, typically in μV
Example	An instantiation of the data received by a model as input, typically denoted by $\mathbf{x}_i$ in the machine learning literature
Trial	A realization of the task under study, e.g. the presentation of one image in a visual ERP paradigm
Window or segment	A group of consecutive EEG samples extracted for further analysis, typically between 0.5 and 30 s
Epoch	A window extracted around a specific trial

other types of data (e.g. images, text and speech) for which DL has been most successful. Therefore, the architectures and practices that are currently used in DL might not be readily applicable to EEG processing.

1.4. Terminology used in this review

Some terms are sometimes used in the fields of machine learning, deep learning, statistics, EEG and signal processing with different meanings. For example, in machine learning, 'sample' usually refers to one example of the input received by a model, whereas in statistics, it can be used to refer to a group of examples taken from a population. It can also refer to the measure of a single time point in signal processing and EEG. Similarly, in deep learning, the term 'epoch' refers to one pass through the whole training set during training; in EEG, an epoch is instead a grouping of consecutive EEG time points extracted around a specific marker. To avoid the confusion, we include in table 1 definitions for a few terms as used in this review. Figure 1 gives a visual example of what these terms refer to.

1.5. Objectives of the review

This systematic review covers the current state-of-the-art in DL-based EEG processing by analyzing a large number of recent publications. It provides an overview of the field for researchers familiar with traditional EEG processing techniques and who are interested in applying DL to their data. At the same time, it aims to introduce the field applying DL to EEG to DL researchers interested in expanding the types of data they benchmark their algorithms with, or who want to contribute to EEG research. For readers in any of these scenarios, this review also provides detailed methodological information on the various components of a DL-EEG pipeline to inform their own implementation[6]. In addition to reporting trends and highlighting interesting approaches, we distill our analysis into a few recommendations in the hope of fostering reproducible and efficient research in the field.

1.6. Organization of the review

The review is organized as follows: section 1 briefly introduces key concepts in EEG and DL, and details the aims of the review; section 2 describes how the systematic review was conducted, and how the studies were selected, assessed and analyzed; section 3 focuses on the most important characteristics of the studies selected and describes trends and promising approaches; section 4 discusses critical topics and challenges in DL-EEG, and provides recommendations for future studies; and section 5 concludes by suggesting future avenues of research in DL-EEG. Finally, supplementary material containing our full data collection table, as well as the code used to produce the graphs, tables and results reported in this review, are made available online.

[6] Additional information with more fine-grained data can be found in our data items table available at http://dl-eeg.com

How to use this review

We advise readers interested in a specific application domain of EEG to use the review in the following way:

1. Read or glance over the main result sections and corresponding discussion sections covering general data items. This should give the reader a broad overview of the current practices and design choices used in the field of DL-EEG.
2. If the reader is interested in a specific application (e.g. brain–computer interfacing) or in a specific type of architecture (e.g. CNNs), identify relevant references in table 4.
3. Consult the detailed summary of the relevant references—which includes the data items introduced in table 3 as well as many more (e.g. detailed preprocessing and feature extraction methodology, software implementation, values of specific hyperparameters, etc)—contained in the data items spreadsheet available online at http://dl-eeg.com.
4. Check the data items spreadsheet for regular updates (including additional studies) and the online repository for an updated version of the figures included in this review. We aim to provide readers with evolving and up-to-date information on various domains and architectures; therefore the table will remain open for external contributions from the community and authors of DL-EEG studies.
5. Use our checklist provided in appendix B to ensure that you include all the relevant information in your future DL-EEG publications.

Table 2. Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
• Training of one or multiple deep learning architecture(s) to process non-invasive EEG data	• Studies focusing solely on invasive EEG (e.g. electrocorticography (ECoG) and intracortical EEG) or magnetoencephalography (MEG) • Papers focusing solely on software tools • Review articles

2. Methods

English journal and conference papers, as well as electronic preprints, published between January 2010 and July 2018, were chosen as the target of this review. PubMed, Google Scholar and arXiv were queried[7] to collect an initial list of papers containing specific search terms in their title or abstract[8]. Additional papers were identified by scanning the reference sections of these papers. The databases were queried for the last time on July 2, 2018.

The following search terms were used to query the databases: 1. EEG, 2. electroencephalogra*, 3. deep learning, 4. representation learning, 5. neural network*, 6. convolutional neural network*, 7. ConvNet, 8. CNN, 9. recurrent neural network*, 10. RNN, 11. long short-term memory, 12. LSTM, 13. generative adversarial network*, 14. GAN, 15. autoencoder, 16. restricted boltzmann machine*, 17. deep belief network* and 18. DBN. The search terms were further combined with logical operators in the following way: (1 OR 2) AND (3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12 OR 13 OR 14 OR 15 OR 16 OR 17 OR 18). The papers were then included or excluded based on the criteria listed in table 2.

To assess the eligibility of the selected papers, the titles were read first. If the title did not clearly indicate whether the inclusion and exclusion criteria were met, the abstract was read as well. Finally, when reading the full text during the data collection process, papers that were found to be misaligned with the criteria were rejected.

Non-peer reviewed papers, such as arXiv electronic preprints[9], are a valuable source of state-of-the-art information as their release cycle is typically shorter than that of peer-reviewed publications. Moreover, unconventional research ideas are more likely to be shared in such repositories, which improves the diversity of the reviewed work and reduces the bias possibly introduced by the peer-review process [140]. Therefore, non-peer reviewed preprints were also included in our review. However, whenever a peer-reviewed publication followed a preprint submission, the peer-reviewed version was used instead.

A data extraction table was designed containing different data items relevant to our research questions, based on previous reviews with similar scopes and the authors' prior knowledge of the field. Following a first inspection of the papers with the data extraction sheet, data items were added, removed and refined. Each paper was initially reviewed by a single author, and then reviewed by a second if needed. For each article selected, around 70 data items were extracted covering five categories: origin of the article, rationale, data used, EEG processing methodology, DL methodology and reported results. Table 3 lists and defines the different items included in each of these categories. We make this data extraction table openly available for interested readers to reproduce our results and dive deeper into the data collected. We also invite authors of published work in the field of DL and EEG to contribute to the table by verifying its content or by adding their articles to it.

The first category covers the origin of the article, that is whether it comes from a journal, a conference publication or a preprint repository, as well as the country of the first author's affiliation. This gives a quick overview of the types of publication included in this review and of the main actors in the field. Second, the rationale category focuses on the domains of application of the selected studies. This is valuable information to understand the extent of the research in the field, and also enables us to identify trends across and within domains in our analysis. Third, the data category includes all relevant information on the data used by the selected papers. This comprises both the origin of the data and the data collection parameters, in addition to the amount of data that was available in each study. Through this section, we aim to clarify the data requirements for using DL on EEG. Fourth, the EEG processing parameters category highlights the typical transformations required to apply DL to EEG, and covers preprocessing steps, artifact handling methodology, as well as feature extraction. Fifth, details of the DL methodology, including architecture design, training procedures and inspection methods, are reported to guide the interested reader through state-of-the-art techniques. Sixth, the reported results category reviews the results of the selected articles, as well as how they were reported, and aims to clarify how DL fares against traditional processing pipelines performance-wise. Finally, the reprodu-cibility of the selected articles is quantified by looking at the availability of the data and code. The results of this section support the critical component of our discussion.

3. Results

The database queries yielded 553 different results that matched the search terms (see figure 2). 49 additional papers were then identified using the reference sections of the initial papers. Based on our inclusion and exclusion criteria, 448 papers were excluded. One additional paper was excluded since it had been retracted. Therefore, 154 papers were selected for inclusion in the analysis.

[7] The queries used for each database are available at http://dl-eeg.com

[8] Since the Google Scholar search engine only allows searching full text or titles, and not titles and abstracts, the query was performed using the flag *allintitle* to search titles only. On arXiv and PubMed, however, both abstracts and titles were queried.

[9] https://arxiv.org/

Table 3. Data items extracted for each article selected.

Category	Data item	Description
Origin of article	Type of publication	Whether the study was published as a journal article, a conference paper or in an electronic preprint repository
	Venue	Publishing venue, such as the name of a journal or conference
	Country of first author affiliation	Location of the affiliated university, institute or research body of the first author
Study rationale	Domain of application	Primary area of application of the selected study. In the case of multiple domains of application, the domain that was the focus of the study was retained
Data	Quantity of data	Quantity of data used in the analysis, reported both in total number of samples and total minutes of recording
	Hardware	Vendor and model of the EEG recording device used
	Number of channels	Number of EEG channels used in the analysis. May differ from the number of recorded channels
	Sampling rate	Sampling rate (reported in Hertz) used during the EEG acquisition
	Subjects	Number of subjects used in the analysis. May differ from the number of recorded subjects
	Data split and cross-validation	Percentage of data used for training, validation, and test, along with the cross-validation technique used, if any
	Data augmentation	Data augmentation technique used, if any, to generate new examples
EEG processing	Preprocessing	Set of manipulation steps applied to the raw data to prepare it for use by the architecture or for feature extraction
	Artifact handling	Whether a method for cleaning artifacts was applied
	Features	Output of the feature extraction procedure, which aims to better represent the information of interest contained in the preprocessed data
Deep learning methodology	Architecture	Structure of the neural network in terms of types of layers (e.g. fully-connected, convolutional)
	Number of layers	Measure of architecture depth
	EEG-specific design choices	Particular architecture choices made with the aim of processing EEG data specifically
	Training procedure	Method applied to train the neural network (e.g. standard optimization, unsupervised pre-training followed by supervised fine-tuning, etc)
	Regularization	Constraint on the hypothesis class intended to improve a learning algorithm generalization performance (e.g. weight decay, dropout)
	Optimization	Parameter update rule
	Hyperparameter search	Whether a specific method was employed in order to tune the hyperparameter set
	Subject handling	Intra- versus inter-subject analysis
	Inspection of trained models	Method used to inspect a trained DL model
Results	Type of baseline	Whether the study included baseline models that used traditional processing pipelines, DL baseline models, or a combination of the two
	Performance metrics	Metrics used by the study to report performance (e.g. accuracy, f1-score, etc)
	Validation procedure	Methodology used to validate the performance of the trained models, including cross-validation and data split
	Statistical testing	Types of statistical tests used to assess the performance of the trained models
	Comparison of results	Reported results of the study, both for the trained DL models and for the baseline models
Reproducibility	Dataset	Whether the data used for the experiment comes from private recordings or from a publicly available dataset
	Code	Whether the code used for the experiment is available online or not, and if so, where

3.1. Origin of the selected studies

Our search methodology returned 51 journal papers, 61 conference and workshop papers and 42 preprints that met our criteria. A total of 28 journal and conference papers had initially been made available as preprints on arXiv or bioRxiv. Popular journals included *Neurocomputing*, *Journal of Neural Engineering* and *Biomedical Signal Processing and Control*, each with three publications contained in our selected studies. We also looked at the location of the first author's affiliation to get a sense of the geographical distribution of research on DL-EEG. We found that most contributions came from the USA, China and Australia (see figure 3).

3.2. Domains

The selected studies applied DL to EEG in various ways (see figure 4 and table 4). Most studies (86%) focused on using DL for the classification of EEG data, most notably for sleep staging, seizure detection and prediction, brain–computer interfaces (BCIs), as well as for cognitive and affective monitoring. Around 9% of the studies focused instead on the improvement of processing tools, such as learning features from EEG, handling artifacts, or visualizing trained models. The remaining papers (5%) explored ways of generating data from EEG, e.g. augmenting data, or generating images conditioned on EEG.

Despite the absolute number of DL-EEG publications being relatively small as compared to other DL applications such as computer vision [98], there is clearly a growing interest in the field. Figure 5 shows the growth of the DL-EEG literature since 2010. The first seven months of 2018 alone count more publications than 2010–2016 combined, hence the relevance of this review. It is, however, still too early to conclude on trends concerning the application domains, given the relatively small number of publications to date.

3.3. Data

The availability of large datasets containing unprecedented numbers of examples is often mentioned as one of the main enablers of deep learning research in the early 2010s [61]. It is thus crucial to understand what the equivalent is in EEG research, given the relatively high cost of collecting EEG data. Given the high dimensionality of EEG signals [116], one would assume that a considerable amount of data is required. Although our analysis cannot answer that question fully, we seek to cover as many dimensions of the answer as possible to give the reader a complete view of what has been done so far.

3.3.1. Quantity of data. We make use of two different measures to report the amount of data used in the reviewed studies: (1) the number of examples available to the deep learning network and (2) the total duration of the EEG recordings used in the study, in minutes. Both measures include the EEG data used across training, validation and test phases. For an in-depth analysis of the amount of data, please see the data items table which contains more detailed information.

The left column of figure 6 shows the amount of EEG data, in minutes, used in the analysis of each study, including training, validation and/or testing. Therefore, the time reported here does not necessarily correspond to the total recording time of the experiment(s). For example, many studies recorded a baseline at the beginning and/or at the end but did not use it in their analysis. Moreover, some studies recorded more classes than they used in their analysis. Also, some studies used sub-windows of recorded epochs (e.g. in a motor imagery BCI, using 3 s of a 7 s epoch). The amount of data in minutes used across the studies ranges from 2 up to 4800000 (mean = 62602; median = 360).

The center column of figure 6 shows the amount of examples available to the models, either for training, validation or test. This number presents a relevant variability as some studies used a sliding window with a significant overlap generating many examples (e.g. 250 ms windows with 234 ms overlap, therefore generating 4050000 examples from 1080 min of EEG data [172]), while some other studies used very long windows generating very few examples (e.g. 15 min windows with no overlap, therefore generating 62 examples from 930 min of EEG data [56]). The wide range of windowing approaches (see section 3.3.4) indicates that a better understanding of its impact is still required. The number of examples used ranged from 62 up to 9750000 (mean = 251532; median = 14000).

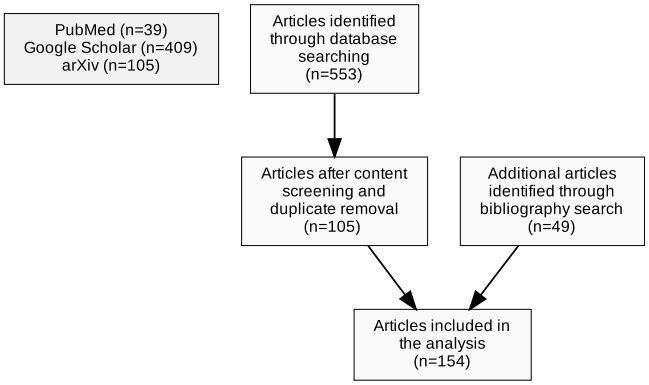

Figure 2. Selection process for the papers.

The right column of figure 6 shows the ratio between the amount of data in minutes and the number of examples. This ratio was never mentioned specifically in the papers reviewed but we nonetheless wanted to see if there were any trends or standards across domains and we found that in sleep studies for example, this ratio tends to be of two as most people are using 30 s non-overlapping windows. Brain–computer interfacing is seeing the most sparsity perhaps indicating a lack of best practices for sliding windows. It is important to note that the BCI field is also the one in which the exact relevant time measures were hardest to obtain since most of the recorded data is not used (e.g. baseline, in-between epochs). Therefore, some of the sparsity on the graph could come from us trying our best to understand and calculate the amount of data used (i.e. seen by the model). Obviously, in the following categories: generation of data, improvement of processing tools and others, this ratio has little to no value as the trends would be difficult to interpret.

The amount of data across different domains varies significantly. In domains like sleep and epilepsy, EEG recordings last many hours (e.g. a full night), but in domains like affective and cognitive monitoring, the data usually comes from lab experiments on the scale of a few hours or even a few minutes.

3.3.2. Subjects. Often correlated with the amount of data, the number of subjects also varies significantly across studies (see figure 7). Half of the datasets used in the selected studies contained fewer than 13 subjects. Six studies, in particular, used datasets with a much greater number of subjects: [145, 166, 178, 207] all used datasets with at least 250 subjects, while [25] and [57] used datasets with 10000 and 16000 subjects, respectively. As explained in section 3.7.4, the untapped

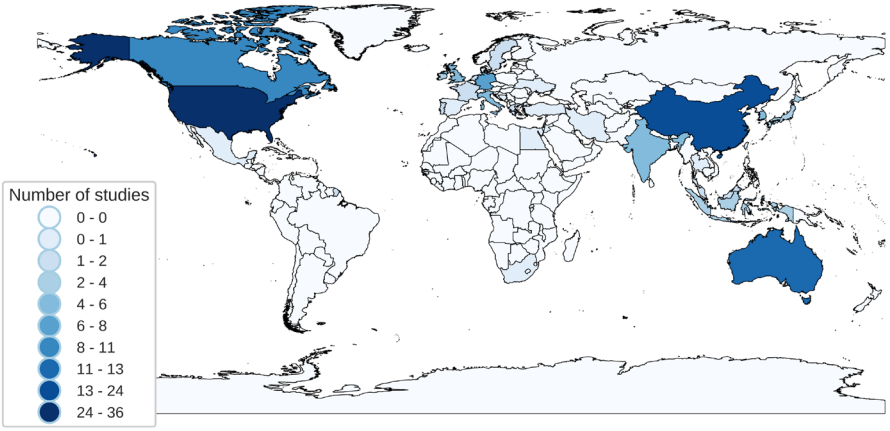

Figure 3. Countries of first author affiliations.

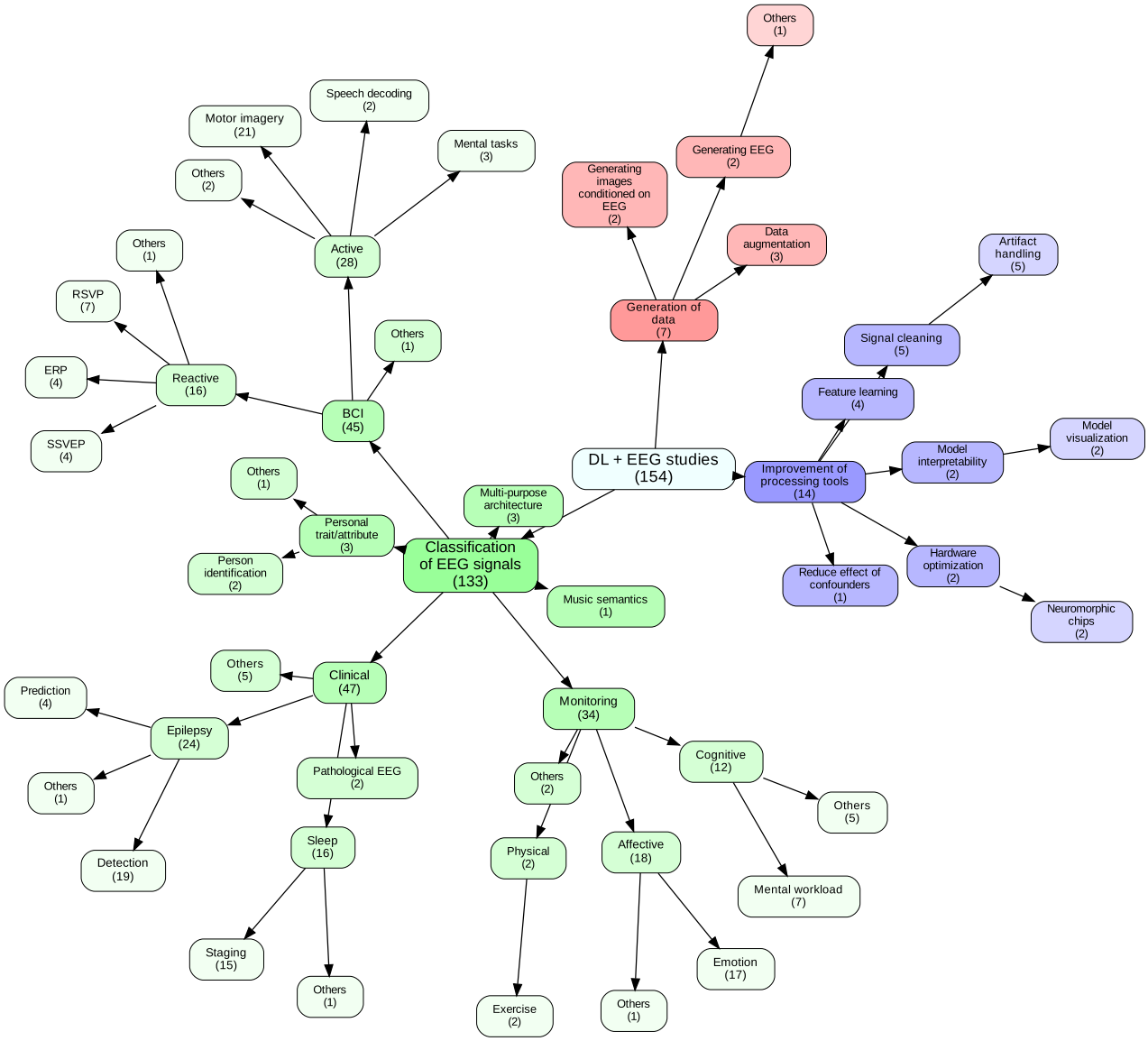

Figure 4. Focus of the studies. The number of papers that fit in a category is showed in brackets for each category. Studies that covered more than one topic were categorized based on their main focus.

Table 4. Categorization of the selected studies according to their application domain and DL architecture. Domains are divided into four levels, as described in figure 4.

Domain 1	Domain 2	Domain 3	Domain 4	AE	CNN	CNN + RNN	DBN	FC	GAN	N/M	Other	RBM	RNN
Classification of EEG signals	BCI	Active	Grasp and lift										[8]
			Mental tasks					[139]					[77, 146]
			Motor imagery	[104, 245]	[43, 50, 115, 161, 162, 167, 189, 191, 223]	[235]	[9]	[34, 67, 118, 124]			[132, 243]		[29, 241]
			Slow cortical potentials								[44]		
			Speech decoding				[180]					[185]	
		Active & Reactive	MI & ERP		[97]								
		Reactive	ERP		[17, 31, 211, 230]								
			Heard speech decoding										[125]
			RSVP		[31, 71, 120, 144, 172, 232]	[179]							
			SSVEP	[147]	[12, 93, 214]								
	Clinical	Alzheimer's disease			[126]								
		Anomaly detection					[218]						
		Dementia		[127]									
		Epilepsy	Detection	[57, 231]	[2, 72, 134, 138, 194, 204]	[58, 59, 171, 207]	[203]	[135]		[192]			[4, 85, 128, 190]
			Event annotation							[224]			
			Prediction		[141, 200]						[199]		[202]
		Ischemic stroke			[56]								
		Pathological EEG			[166]	[157]							
		Schizophrenia	Detection		[35]								
		Sleep	Abnormality detection		[158]								
			Staging	[96, 198]	[33, 121, 145, 150, 178, 201, 209, 219]	[25, 186]	[96]				[55]		[45]
	Monitoring	Affective	Bullying incidents		[14]								
			Emotion	[19, 88, 114, 220]	[112, 113]		[106, 247, 248]	[49, 94, 193]		[123]		[51]	[6, 111, 239]
		Cognitive	Drowsiness									[69]	
			Engagement									[103]	

(*Continued*)

Table 4. (*Continued*)

Domain 1	Domain 2	Domain 3	Domain 4	AE	CNN	CNN + RNN	DBN	FC	GAN	N/M	Other	RBM	RNN
			Eyes closed/open				[129]						
			Fatigue		[70]								
			Mental workload	[226, 227]	[7, 237]	[80, 92]							[81]
			Mental workload & fatigue				[228]						
		Cognitive versus affective					[15]						
		Music semantics			[181]								
		Physical	Exercise		[53]					[63]			
	Multi-purpose architecture			[101]	[244]						[41]		
	Music semantics				[153]								
	Personal trait/ attribute	Person identification											[240, 242]
		Sex			[207]								
Generation of data	Data augmentation				[170, 212, 238]								
	Generating EEG								[75]				
		Spatial upsampling							[40]				
	Generating images conditioned on EEG								[102]		[142]		
Improvement of processing tools	Feature learning			[105, 182, 215]	[16]								
	Hardware optimization	Neuromorphic chips			[133]					[225]			
	Model interpretability	Model visualization			[76]					[183]			
	Reduce effect of confounders												[217]
	Signal cleaning	Artifact handling		[221, 222]	[213]			[46]					[143]

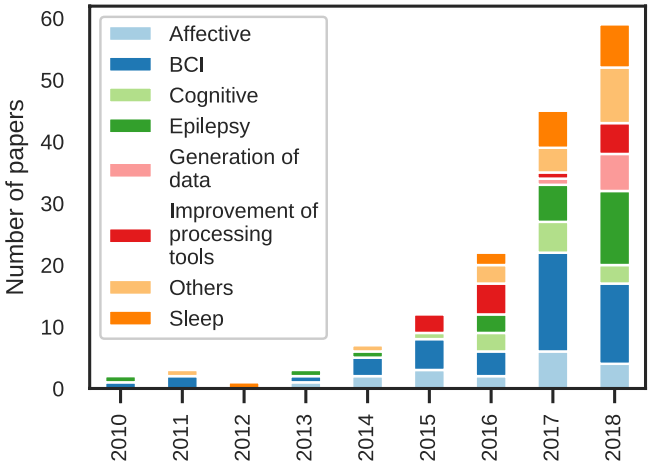

Figure 5. Number of publications per domain per year. To simplify the figure, some of the categories defined in figure 4 have been grouped together.

potential of DL-EEG might reside in combining data coming from many different subjects and/or datasets to train a model that captures common underlying features and generalizes better. In [222], for example, the authors trained their model using an existing public dataset and also recorded their own EEG data to test the generalization on new subjects. In [211], an increase in performance was observed when using more subjects during training before testing on new subjects. The authors tested using from 1 to 30 subjects with a leave-one-subject-out cross-validation scheme, and reported an increase in performance with noticeable diminishing returns above 15 subjects.

3.3.3. Recording parameters. As shown later in section 3.8, 42% of reported results came from private recordings. We look at the type of EEG device that was used by the selected studies to collect their data, and additionally highlight low-cost, often called 'consumer' EEG devices, as compared to traditional 'research' or 'medical' EEG devices (see figure 8(a)). We loosely defined low-cost EEG devices as devices under the USD 1000 threshold (excluding software, licenses and accessories). Among these devices, the Emotiv EPOC was used the most, followed by the OpenBCI, Muse and Neurosky devices. As for the research grade EEG devices, the BioSemi ActiveTwo was used the most, followed by BrainProducts devices.

The EEG data used in the selected studies was recorded with 1–256 electrodes, with half of the studies using between 8 and 62 electrodes (see figure 8(b)). The number of electrodes required for a specific task or analysis is usually arbitrarily defined as no fundamental rules have been established. In most cases, adding electrodes will improve possible analyses by increasing spatial resolution. However, adding an electrode close to other electrodes might not provide significantly different information, while increasing the preparation time and the participant's discomfort and requiring a more costly device. Higher density EEG devices are popular in research but hardly ecological. In [171], the authors explored the impact of the number of channels on the specificity and sensitivity for seizure detection. They showed that increasing the number of channels from 4 up to 22 (including two referential channels) resulted in an increase in sensitivity from 31% to 39% and from 40% to 90% in specificity. They concluded, however, that the position of the referential channels is very important as well, making it difficult to compare across datasets coming from different neurologists and recording sites using different locations for the reference(s) channel(s).

Similarly, in [33], the impact of different electrode configurations was assessed on a sleep staging task. The authors found that increasing the number of electrodes from two to six produced the highest increase in performance, while adding additional sensors, up to 22 in total, also improved the performance but not as much. The placement of the electrodes in a 2-channel montage also impacted the performance, with central and frontal montages leading to better performance than posterior ones on the sleep staging task.

Furthermore, the recording sampling rates varied mostly between 100 and 1000 Hz in the selected studies. As described in section 3.4, however, it is common to decrease the EEG sampling rate before further processing—a process called downsampling, by which a signal is resampled to reduce its dimensionality, often by keeping every other N points. Around 50% of studies used sampling rates of 250 Hz or less and the highest sampling rate used was 5000 Hz [76].

3.3.4. Data augmentation. Data augmentation is a technique by which new data examples are artificially generated from the existing training data. Data augmentation has proven efficient in other fields such as computer vision, where data manipulations including rotations, translations, cropping and flipping can be applied to generate more training examples [147]. Adding more training examples allows the use of more complex models comprising more parameters while reducing overfitting. When done properly, data augmentation increases accuracy and stability, offering a better generalization on new data [234].

Out of the 154 papers reviewed, three papers explicitly explored the impact of data augmentation on DL-EEG ([170, 212, 238]). Interestingly, each one looked at it from the perspective of a different domain: sleep, affective monitoring and BCI. Also, all three are from 2018, perhaps showing an emerging interest in data augmentation. First, in [212], Gaussian noise was added to the training data to obtain new examples. This approach was tested on two different public datasets for emotion classification (SEED [246] and MAHNOB-HCI [177]). They improved their accuracy on the SEED dataset using LeNet ([100]) from 49.6% (without augmentation) to 74.3% (with augmentation), from 34.2% (without) to 75.0% (with) using ResNet ([79]) and from 40.8% (without) to 45.4% (with) on MAHNOB-HCI dataset using ResNet. Their best accuracy was obtained with a standard deviation of 0.2 and by augmenting the data to 30 times its original size. Despite impressive results, it is important to note that they also compared LeNet and ResNet to an SVM which had an accuracy of 74.2% (without) and 73.4% (with) on the SEED dataset. This might indicate that the initial amount of data was insufficient for LeNet or ResNet but adding data

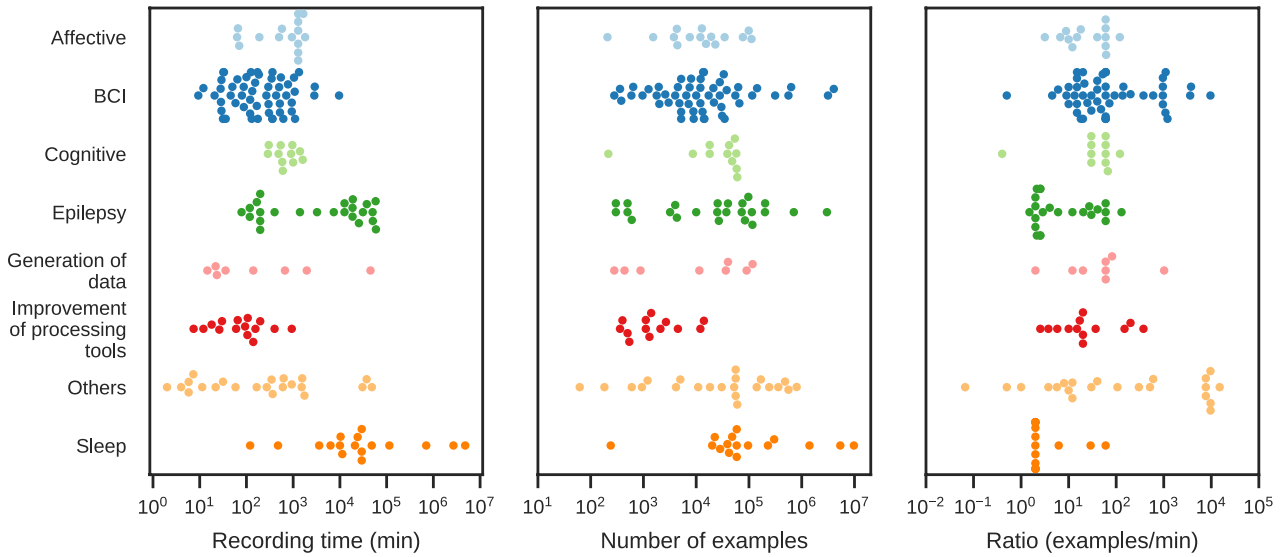

Figure 6. Amount of data used by the selected studies. Each dot represents one dataset. The left column shows the datasets according to the total length of the EEG recordings used, in minutes. The center column shows the number of examples that were extracted from the available EEG recordings. The right column presents the ratio of number of examples to minutes of EEG recording.

clearly helped bring the performance up to par with the SVM. Second, in [238], a conditional deep convolutional generative adversarial network (cDCGAN) was used to generate artificial EEG signals on one of the BCI Competition motor imagery datasets. Using a CNN, it was shown that data augmentation helped improve accuracy from 83% to around 86% to classify motor imagery. In [170], the authors explicitly targeted the class imbalance problem of under-represented sleep stages by generating Fourier transform (FT) surrogates of raw EEG data on the CAPSLPDB dataset. They improved their accuracy up to 24% on some classes.

An additional 30 papers explicitly used data augmentation in one form or another but only a handful investigated the impact it has on performance. In [16, 92], noise was added to 2D feature images, although it did not improve results in [16]. In [85], artifacts such as eye blinks and muscle activity, as well as Gaussian white noise, were used to augment the data and improve robustness. In [228] and [227], Gaussian noise was added to the input feature vector. This approach increased the accuracy of the SDAE model from around 76.5% (without augmentation) to 85.5% (with).

Multiple studies also used overlapping windows as a way to augment their data, although many did not explicitly frame this as data augmentation. In [134, 204], overlapping windows were explicitly used as a data augmentation technique. In [93], different shift lengths between overlapping windows (from 10 ms to 60 ms out of a 2 s window) were compared, showing that by generating more training samples with smaller shifts, performance improved significantly. In [167], the concept of overlapping windows was pushed even further: (1) redundant computations due to EEG samples being in more than one window were simplified thanks to 'cropped training', which ensured these computations were only done once, thereby speeding up training and (2) the fact that overlapping windows share information was used to design an additional term to the cost function, which further regularizes the models by penalizing decisions that are not the same while being close in time.

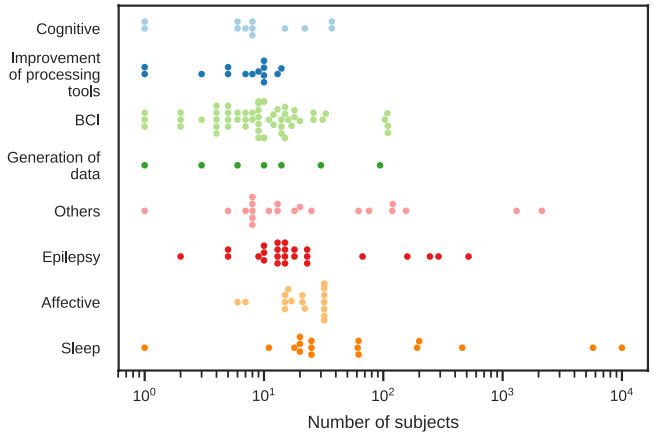

Figure 7. Number of subjects per domain in datasets. Each point represents one dataset used by one of the selected studies.

Other procedures used the inherent spatial and temporal characteristics of EEG to augment their data. In [41], the authors doubled their data by swapping the right and left side electrodes, claiming that as the task was a symmetrical problem, which side of the brain expresses the response would not affect classification. In [19], the authors augmented their multimodal (EEG and EMG) data by duplicating samples and keeping the values from one modality only, while setting the other modality values to 0 and vice-versa. In [49], the authors made use of the data that is usually thrown away when downsampling EEG in the preprocessing stage. It is common to downsample a signal acquired at higher sampling rate to 256 Hz or less. In their case, they reused the data thrown away during that step as new samples: a downsampling by a factor of N would therefore allow an augmentation of N times.

Finally, classification of rare events where the number of available samples are orders of magnitude smaller than

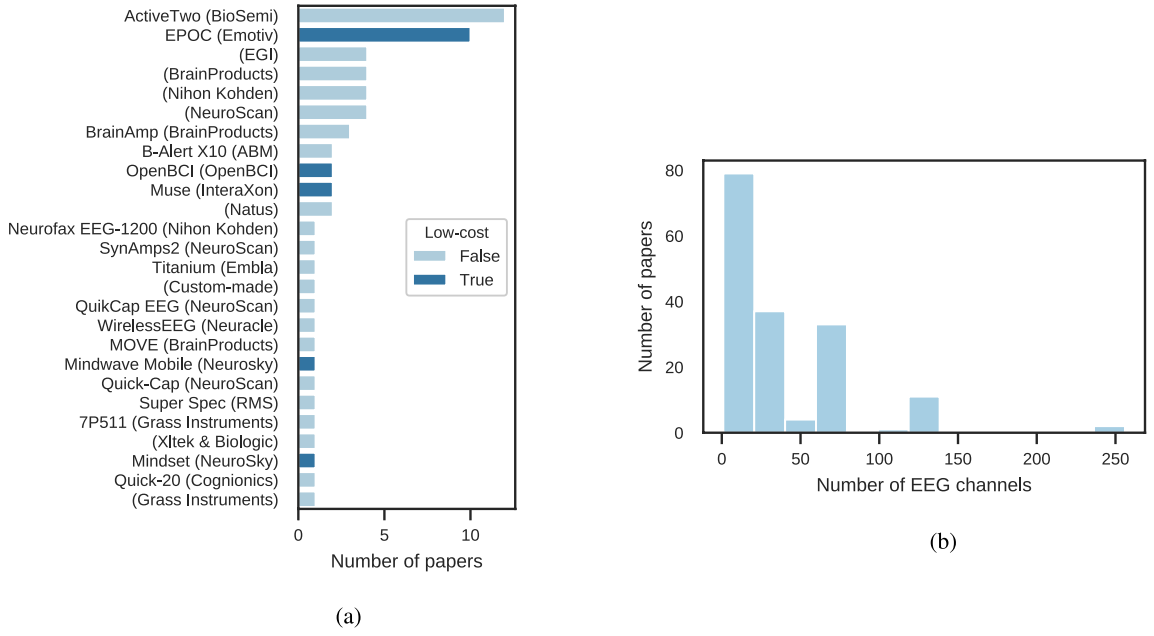

Figure 8. Hardware characteristics of the EEG devices used to collect the data in the selected studies. (a) EEG hardware used in the studies. The device name is followed by the manufacturer's name in parentheses. Low-cost devices (defined as devices below $1000 excluding software, licenses and accessories) are indicated by a different color. (b) Distribution of the number of EEG channels.

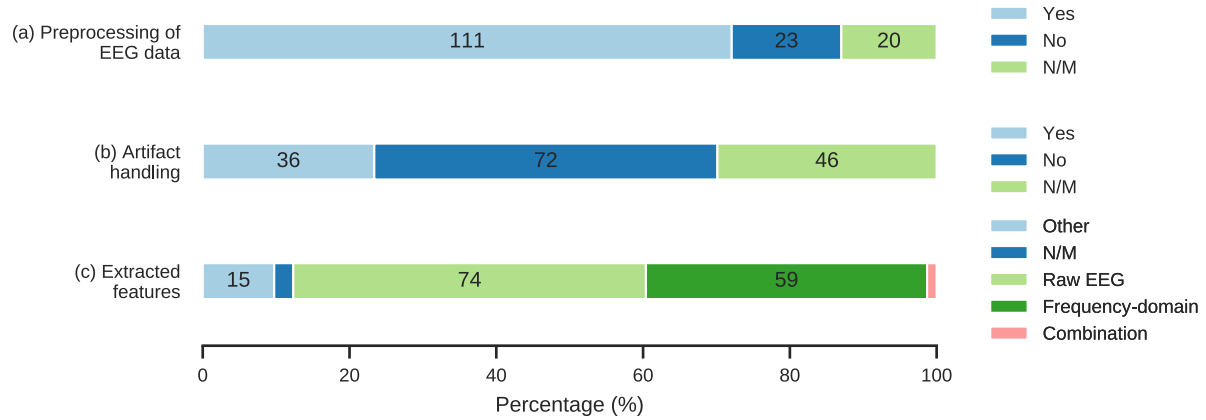

Figure 9. EEG processing choices. (a) Number of studies that used preprocessing steps, such as filtering, (b) number of studies that included, rejected or corrected artifacts in their data and (c) types of features that were used as input to the proposed models.

their counterpart classes [170] is another motivation for data augmentation. In EEG classification, epileptic seizures or transitional sleep stages (e.g. S1 and S3) often lead to such unbalanced classes. In [209], the class imbalance problem was addressed by randomly balancing all classes while sampling for each training epoch. Similarly, in [33], balanced accuracy was maximized by using a balanced sampling strategy. In [202], EEG segments from the interictal class were split into smaller subgroups of equal size to the preictal class. In [178], cost-sensitive learning and oversampling were used to solve the class imbalance problem for sleep staging but the overall performance using these approaches did not improve. In [158], the authors randomly replicated subjects from the minority class to balance classes. Similarly, in [45, 46, 120, 186], oversampling of the minority class was used to balance classes. Conversely, in [172, 194], the majority class was subsampled. In [200], an overlapping window with a subject-specific overlap was used to match classes. Similar work by the same group [199] showed that when training a GAN on individual subjects, augmenting data with an overlapping window increased accuracy from 60.91% to 74.33%. For more on imbalanced learning, we refer the interested reader to [173].

3.4. EEG processing

One of the oft-claimed motivation for using deep learning on EEG processing is automatic feature learning [12, 53, 77, 85, 125, 145, 232]. This can be explained by the fact that feature engineering is a time-consuming task [109]. Additionally, preprocessing and cleaning EEG signals from artifacts is a demanding step of the usual EEG processing pipeline. Hence, in this section, we look at aspects related to data preparation, such as preprocessing, artifact handling and feature extraction. This analysis is critical to clarify what level of preprocessing

EEG data requires to be successfully used with deep neural networks.

3.4.1. Preprocessing. Preprocessing EEG data usually comprises a few general steps, such as downsampling, band-pass filtering, and windowing. Throughout the process of reviewing papers, we found that a different number of preprocessing steps were employed in the studies. In [80], it is mentioned that 'a substantial amount of preprocessing was required' for assessing cognitive workload using DL. More specifically, it was necessary to trim the EEG trials, downsample the data to 512 Hz and 64 electrodes, identify and interpolate bad channels, calculate the average reference, remove line noise, and high-pass filter the data starting at 1 Hz. On the other hand, Stober *et al* [182] applied a single preprocessing step by removing the bad channels for each subject. In studies focusing on emotion recognition using the DEAP dataset [91], the same preprocessing methodology proposed by the researchers that collected the dataset was typically used, i.e. re-referencing to the common average, downsampling to 256 Hz, and high-pass filtering at 2 Hz.

We separated the papers into three categories based on whether or not they used preprocessing steps: 'Yes', in cases where preprocessing was employed; 'No', when the authors explicitly mentioned that no preprocessing was necessary; and not mentioned ('N/M') when no information was provided. The results are shown in figure 9.

A considerable proportion of the reviewed articles (72%) employed at least one preprocessing method such as downsampling or re-referencing. This result is not surprising, as applications of DNNs to other domains, such as computer vision, usually require some kind of preprocessing like cropping and normalization as well.

3.4.2. Artifact handling. Artifact handling techniques are used to remove specific types of noise, such as ocular and muscular artifacts [205]. As emphasized in [222], removal of artifacts may be crucial for achieving good EEG decoding performance. Adding this to the fact that cleaning EEG signals might be a time-consuming process, some studies attempted to apply only minimal preprocessing such as removing bad channels and leave the burden of learning from a potentially noisy signal on the neural network [182]. With that in mind, we decided to look at artifact handling separately.

Artifact removal techniques usually require the intervention of a human expert [131]. Different techniques leverage human knowledge to different extents, and might fully rely on an expert, as in the case of visual inspection, or require prior knowledge to simply tune a hyperparameter, as in the case of wavelet-enhanced independent component analysis (wICA) [30]. Among the studies which handled artifacts, a myriad of techniques were applied. Some studies employed methods which rely on human knowledge such as amplitude thresholding [125], manual identification of high-variance segments [80], and handling EEG blinking-related noise based on high-amplitude EOG segments [120]. Moreover, in [53, 144, 146, 185, 226, 227], independent component analysis (ICA) was used to separate ocular components from EEG data [119].

In order to investigate the necessity of removing artifacts from EEG when using deep neural networks, we split the selected papers into three categories, in a similar way to the preprocessing analysis (see figure 9). Almost half the papers (47%) did not use artifact handling methods, while 23% did. Additionally, 30% of the studies did not mention whether artifact handling was necessary to achieve their results. Given those results, we are encouraged to believe that using DNNs on EEG might be a way to avoid the explicit artifact removal step of the classical EEG processing pipeline without harming task performance.

3.4.3. Features. Feature engineering is one of the most demanding steps of the traditional EEG processing pipeline [109] and the main goal of many papers considered in this review [12, 53, 77, 85, 125, 145, 232] is to get rid of this step by employing deep neural networks for automatic feature learning. This aspect appears to be of interest to researchers in the field since its early stages, as indicated by the work of Wulsin *et al* [218], which, in 2011, compared the performance of deep belief networks (DBNs) on classification and anomaly detection tasks using both raw EEG and features as inputs. More recently, studies such as [75, 183] achieved promising results without the need to extract features.

On the other hand, a considerable proportion of the reviewed papers used hand-engineered features as the input to their deep neural networks. In [193], for example, authors used a time-frequency domain representation of EEG obtained via the short-time Fourier transform (STFT) for detecting binary user-preference (*like* versus *dislike*). Similarly, Truong *et al* [200], used the STFT as a 2-dimensional EEG representation for seizure prediction using CNNs. In [237], EEG frequency-domain information was also used. Widely adopted by the EEG community, the power spectral density (PSD) of classical frequency bands from around 1 Hz to 40 Hz were used as features. Specifically, authors selected the delta (1–4 Hz), theta (5–8 Hz), alpha (9–13 Hz), lower beta (14–16 Hz), higher beta (17–30 Hz), and gamma (31–40 Hz) bands for mental workload state recognition. Moreover, other studies employed a combination of features, for instance [56], which used PSD features, as well as entropy, kurtosis, fractal component, among others, as input of the proposed CNN for ischemic stroke detection.

Given that the majority of EEG features are obtained in the frequency-domain, our analysis consisted in separating the reviewed articles into four categories according to the respective input type. Namely, the categories were: 'Raw EEG' (which includes EEG time series that have been preprocessed, e.g. filtered or artifact-corrected), 'Frequency-domain', 'Combination' (in case more than one type of feature was used), and 'Other' (for papers using neither raw EEG nor frequency-domain features). Studies that did not specify the type of input were assigned to the category 'N/M' (not mentioned). Notice that, here, we use 'feature' and 'input type' interchangeably.

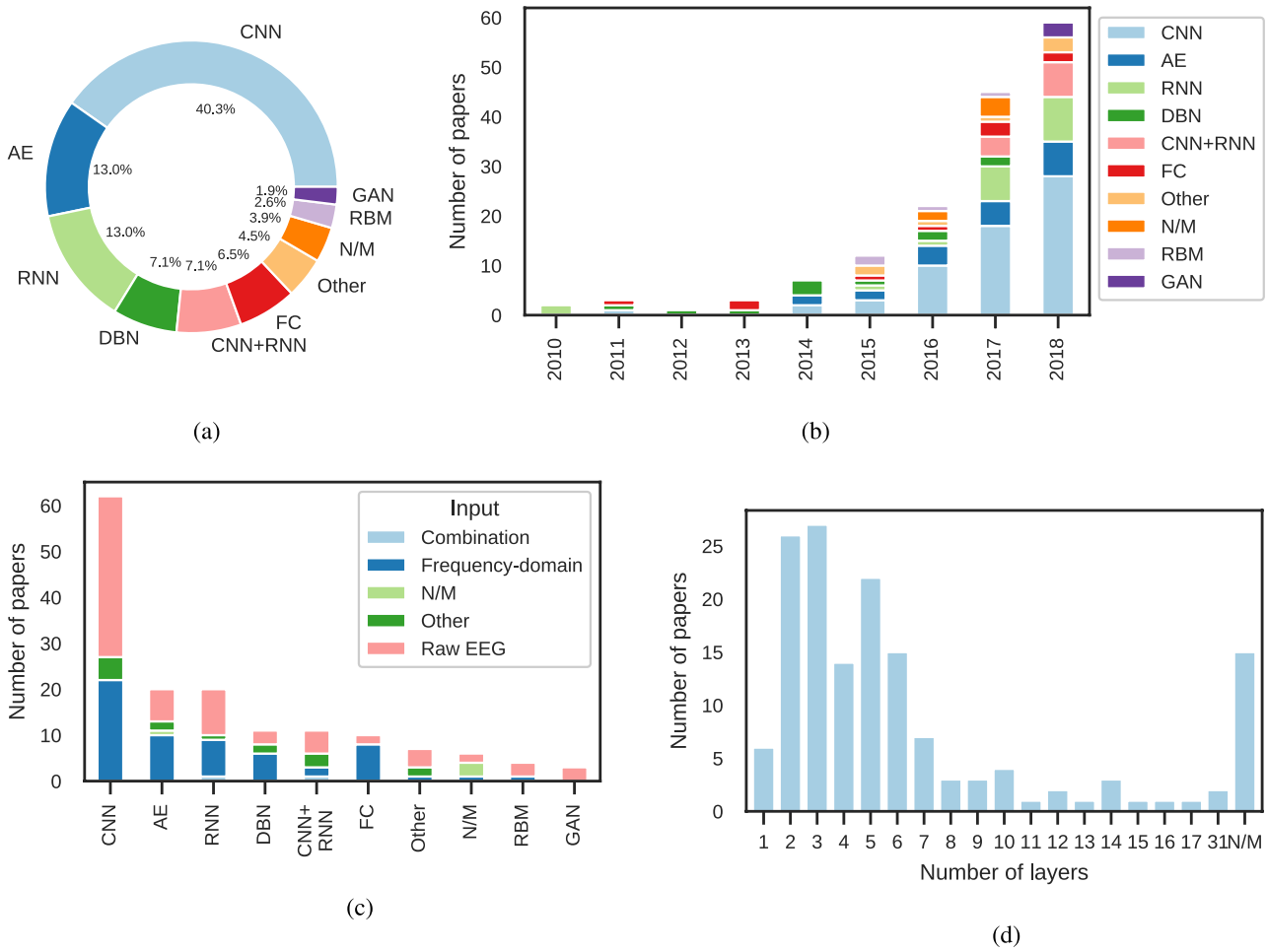

Figure 10. Deep learning architectures used in the selected studies. 'N/M' stands for 'Not mentioned' and accounts for papers which have not reported the respective deep learning methodology aspect under analysis. (a) Architectures. (b) Distribution of architectures across years. (c) Distribution of input type according to the architecture category. (d) Distribution of number of neural network layers.

Figure 9 presents the result of our analysis. One can observe that 49% of the papers used only raw EEG data as input, whereas 49% used hand-engineered features, from which 38% corresponded to frequency domain-derived features. Finally, 2% did not specify the type of input of their model. According to these results, we find indications that DNNs can be in fact applied to raw EEG data and achieve state-of-the-art results.

3.5. Deep learning methodology

3.5.1. Architecture. A crucial choice in the DL-based EEG processing pipeline is the neural network architecture to be used. In this section, we aim at answering a few questions on this topic, namely: (1) 'What are the most frequently used architectures?', (2) 'How has this changed across years?', (3) 'Is the choice of architecture related to input characteristics?' and (4) 'How deep are the networks used in DL-EEG?'.

To answer the first three questions, we divided and assigned the architectures used in the 154 papers into the following groups: CNNs, RNNs, AEs, restricted Boltzmann machines (RBMs), DBNs, GANs, FC networks, combinations of CNNs and RNNs (CNN + RNN), and 'Others' for any other architecture or combination not included in the aforementioned categories. Figure 10(a) shows the percentage of studies that used the different architectures. 40% of the papers used CNNs, whereas RNNs and AEs were both the architecture choice of about 13% of the works, respectively. Combinations of CNNs and RNNs, on the other hand, were used in 7% of the studies. RBMs and DBNs corresponded together to almost 10% of the architectures. FC neural networks were employed by 6% of the papers. GANs and other architectures appeared in 6% of the considered cases. Notice that 4% of the analyzed papers did not report their choice of architecture.

In figure 10(b), we provide a visualization of the distribution of architecture types across years. Until the end of 2014, DBNs and FC networks comprised the majority of the studies. However, since 2015, CNNs have been the architecture type of choice in most studies. This can be attributed to the their capabilities of end-to-end learning and of exploiting hierarchical structure on the data [196], as well as their success and subsequent popularity on computer vision tasks, such as the ILSVRC 2012 challenge [42]. Interestingly, we also observe that as the number of papers grows, the proportion of studies using CNNs and combinations of recurrent and convolutional layers has been growing steadily. The latter shows that RNNs are increasingly of interest for EEG analysis. On the other

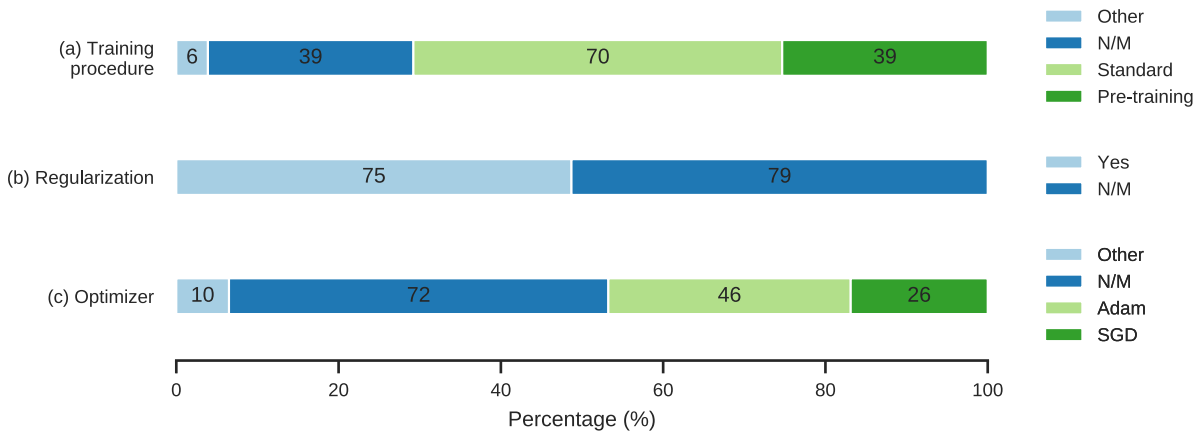

Figure 11. Deep learning methodology choices. (a) Training methodology used in the studies, (b) number of studies that reported the use of regularization methods such as dropout, weight decay, etc and (c) type of optimizer used in the studies.

hand, the use of architectures such as RBMs, DBNs and AEs has been decreasing with time. Commonly, models employing these architectures utilize a two-step training procedure consisting of (1) unsupervised feature learning and (2) training a classifier on top of the learned features. However, we notice that recent studies leverage the hierarchical feature learning capabilities of CNNs to achieve end-to-end supervised feature learning, i.e. training both a feature extractor and a classifier simultaneously.

To complement the previous result, we cross-checked the architecture and input type information provided in figure 9. Results are presented in figure 10(c) and clearly show that CNNs are indeed used more often with raw EEG data as input. This corroborates the idea that researchers employ this architecture with the aim of leveraging the capabilities of deep neural networks to process EEG data in an end-to-end fashion, avoiding the time-consuming task of extracting features. From this figure, one can also notice that some architectures such as deep belief networks are typically used with frequency-domain features as inputs, while GANs, on the other hand, have been only applied to EEG processing using raw data.

Number of layers. Deep neural networks are usually composed of stacks of layers which provide hierarchical processing. Although one might think the use of *deep* neural networks implies the existence of a large number of layers in the architecture, there is no absolute consensus in the literature regarding this definition. Here we investigate this aspect and show that the number of layers is not necessarily large, i.e. larger than three, in many of the considered studies.

In figure 10(d), we show the distribution of the reviewed papers according to the number of layers in the respective architecture. For studies reporting results for different architectures and number of layers, we only considered the highest value. We observed that most of the selected studies (128) utilized architectures with at most 10 layers. A total of 16 articles have not reported the architecture depth. When comparing the distribution of papers according to the architecture depth with architectures commonly used for computer vision applications, such as VGG-16 (16 layers) [175] and ResNet-18 (18 layers) [79], we observe that the current literature on DL-EEG suggests shallower models achieve better performance. The same trend is applicable to other domains such as NLP and speech processing. In unsupervised language modeling, for instance, the GPT-2 model [152] outperformed previous work[10] using architectures with 12–48 layers. Likewise, in automatic speech recognition, the state-of-the-art model[11] on human-to-human communication was achieved with a 30-layer architecture containing residual blocks and recurrent layers [164].

Some studies specifically investigated the effect of increasing the model depth. Zhang *et al* [237] evaluated the performance of models with depth ranging from two to 10 on a mental workload classification task. Architectures with seven layers outperformed both shallower (two and four layers) and deeper (10 layers) models in terms of accuracy, precision, F-measure and G-mean. Moreover, O'Shea *et al* [134] compared the performance of a CNN with six and 11 layers on neonatal seizure detection. Their results show that, in this case, the deeper network presented better area under the receiver operating curve (ROC AUC) (0.971) in comparison to the shallower model, as well as a support vector machine (SVM) (0.965). In [93], the effect of depth on CNN performance was also studied. The authors compared results obtained by a CNN with two and three convolutional layers on the task of classifying SSVEPs under ambulatory conditions. The shallower architecture outperformed the three-layer one in all scenarios considering different amounts of training data. Canonical correlation analysis (CCA) together with a KNN classifier were also evaluated and employed as a baseline method. Interestingly, as the number of training samples increased, the shallower model outperformed the CCA-based baseline. These three examples offer a representative view of the current state of DL-EEG research, namely that it is impossible to conclude that either deeper or shallower models perform better in all contexts. Depending on factors such as the amount of data, the task to be solved, the type of architecture, the hyperparameter tuning strategy, and the available computational resources, shallower or deeper models might work

[10] https://paperswithcode.com/sota/word-level-models-penn-treebank
[11] https://paperswithcode.com/sota/speech-recognition-word-error-rate-on-sw

best. To gain a better idea of what might be preferable to use in a specific case, we invite the reader to identify relevant studies in the data items table and explore their results.

EEG-specific design choices. Particular choices regarding the architecture might enable a model to mimic the process of extracting EEG features. An architecture can also be specifically designed to impose specific properties on the learned representations. This is for instance the case with max-pooling, which is used to produce invariant feature maps to slight translations on the input [61]. In the case of EEG signals, one might be interested in forcing the model to process temporal and spatial information separately in the earlier stages of the network. In [17, 33, 93, 120, 167, 232], one-dimensional convolutions were used in the input layer with the aim of processing either temporal or spatial information independently at this point of the hierarchy. Other studies [186, 243] combined recurrent and convolutional neural networks as an alternative to the previous approach of separating temporal and spatial content. Recurrent models were also applied in cases where it was necessary to capture long-term dependencies from the EEG data [111, 239].

3.5.2. Training. Details regarding the training of the models proposed in the literature are of great importance as different approaches and hyperparameter choices can greatly impact the performance of neural networks. The use of pre-trained models, regularization, and hyperparameter search strategies are examples of aspects we took into account during the review process. We report our main findings in this section.

Training procedure. One of the advantages of applying deep neural networks to EEG processing is the possibility of simultaneously training a feature extractor and a model for executing a downstream task such as classification or regression. However, in some of the reviewed studies [96, 127, 215], these two tasks were executed separately. Usually, the feature learning was done in an unsupervised fashion, with RBMs, DBNs, or AEs. After training those models to provide an appropriate representation of the EEG input signal, the new features were then used as the input for a target task which is, in general, classification. In other cases, pre-trained models were used for a different purpose, such as object recognition, and were fine-tuned on the specific EEG task with the aim of providing a better initialization or regularization effect [108].

In order to investigate the training procedure of the reviewed papers, we classify each one according to the adopted training procedure. Models which have parameters learned without using any kind of pre-training were assigned to the 'Standard' group. The remaining studies, which specified the training procedure, were included in the 'Pre-training' class, in case the parameters were learned in more than one step. Finally, papers employing different methodologies for training, such as co-learning [41], were included in the 'Other' group.

In figure 11(a) we show how the reviewed papers are distributed according to the training procedure. 'N/M' refers to studies which have not reported this aspect. Almost half the papers did not employ any pre-traning strategy, while 25% did. Even though the training strategy is crucial for achieving good performance with deep neural networks, 25% of the selected studies have not explicitly described it in their paper.

Regularization. In the context of our literature review, we define regularization as any constraint on the set of possible functions parametrized by the neural network intended to improve its performance on unseen data during training [61]. The main goal when regularizing a neural network is to control its complexity in order to obtain better generalization performance [24], which can be verified by a decrease on test error in the case of classification problems. There are several ways of regularizing neural networks, and among the most common are weight decay (L2 and L1 regularization) [61], early stopping [151], dropout [187], and label smoothing [188]. Notice that even though the use of pre-trained models as initialization can also be interpreted as a regularizer [108], in this work we decided to include it in the training procedure analysis instead.

As the use of regularization might be fundamental to guarantee a good performance on unseen data during training, we analyzed how many of the reviewed studies explicitly stated that they have employed it in their models. Papers were separated in two groups, namely: 'Yes' in case any kind of regularization was used, and 'N/M' otherwise. In figure 11 we present the proportion of studies in each group.

From figure 11, one can notice that more than half the studies employed at least one regularization method. Furthermore, regularization methods were frequently combined in the reviewed studies. Hefron *et al* [80] employed a combination of dropout, L1- and L2-regularization to learn temporal and frequency representations across different participants. The developed model was trained for recognizing mental workload states elicited by the MATB task [38]. Similarly, Längkvist and Loutfi [96], combined two types of regularization with the aim of developing a model tailored to an automatic sleep stage classification task. Besides L2-regularization, they added a penalty term to encourage weight sparsity, defined as the KL-divergence between the mean activation of each hidden unit over all training examples in a training batch and a hyperparameter ρ.

Optimization. Learning the parameters of a deep neural network is, in practice, an optimization problem. The best way to tackle it is still an open research question in the deep learning literature, as there is often a compromise between finding a good solution in terms of minimizing the cost function and the performance of a local optimum expressed by the generalization gap, i.e. the difference between the training error and the true error estimated on the test set. In this scenario, the choice of a parameter update rule, i.e. the *learning algorithm* or *optimizer*, might be key for achieving good results.

The most commonly used optimizers are reported in figure 11. One surprising finding is that even though the choice of optimizer is a fundamental aspect of the DL-EEG pipeline, 47% of the considered studies did not report which parameter update rule was applied. Moreover, 30% used Adam [90] and 17% Stochastic Gradient Descent [154] (notice that we also

Table 5. Model inspection techniques used by more than one study.

	Citation
Analysis of weights	[32, 41, 95–97, 120, 133, 148, 182, 189, 201, 220, 223, 230, 247]
Analysis of activations	[93, 97, 120, 172, 186, 214, 227, 231]
Input-perturbation network-prediction correlation maps	[17, 76, 166, 167, 211]
Generating input to maximize activation	[16, 158, 178, 207]
Occlusion of input	[33, 102, 194]

refer to the mini-batch case as SGD). 6% of the papers utilized different optimizers, such as RMSprop [197], Adagrad [47], and Adadelta [233].

Another interesting finding the optimizer analysis provided is the steady increase in the use of Adam. Indeed, from 2017 to 2018, the percentage of studies using Adam increased from 28.9% to 54.2%. Adam was proposed as a gradient-based method with the capability of adaptively tuning the learning rate based on estimates of first and second order moments of the gradient. It became very popular in general deep neural networks applications (accumulating approximately 15 000 citations since 2014[12]). Interestingly, we notice a proportional decrease from 2017 to 2018 of the number of papers which did not report the optimizer utilized.

Hyperparameter search. From a practical point-of-view, tuning the hyperparameters of a learning algorithm often takes up a great part of the time spent during training. GANs, for instance, are known to be sensitive to the choices of optimizer and architecture hyperparameters [66, 110]. In order to minimize the amount of time spent finding an appropriate set of hyperparameters, several methods have been proposed in the literature. Examples of commonly applied methods are grid search [20] and Bayesian optimization [176]. Grid search consists in determining a range of values for each parameter to be tuned, choosing values in this range, and evaluating the model, usually in a validation set considering all combinations. One of the advantages of grid search is that it is highly parallelizable, as each set of hyperparameter is independent of the other. Bayesian optimization, in turn, defines a posterior distribution over the hyperparameters space and iteratively updates its values according to the performance obtained by the model with a hyperparameter set corresponding to the expected posterior.

Given the importance of finding a good set of hyperparameters and the difficulty of achieving this in general, we calculate the percentage of papers that employed some search method for tuning their models and optimizers, as well as the amount of articles that have not included any information regarding this aspect. Results indicate that 80% of the reviewed papers have not mentioned the use of hyperparameters search strategies. It is important to highlight that among those articles, it is not clear how many have not done any tuning at all and how many have just not considered to include this information in the paper. From the 20% that declared to have searched for an appropriate set of hyperparameters, some have manually done this by trial and error (e.g. [2, 45, 145, 202]), while others employed grid search (e.g. [12, 46, 96, 112, 220, 226, 227]), and a few used other strategies such as Bayesian methods (e.g. [170, 181, 182]).

[12] Google scholar query run on 30/11/2018.

3.6. Inspection of trained models

In this section, we review if, and how, studies have inspected their proposed models. Out of the selected studies, 27% reported inspecting their models. Two studies focused more specifically on the question of model inspection in the context of DL and EEG [53, 76]. See table 5 for a list of the different techniques that were used by more than one study. For a general review of DL model inspection techniques, see [84].

The most frequent model inspection techniques involved the analysis of the trained model's weights [41, 96, 97, 120, 133, 147, 182, 189, 201, 220, 223, 230, 246]. This often requires focusing on the weights of the first layer only, as their interpretation in regard to the input data is straightforward. Indeed, the absolute value of a weight represents the strength with which the corresponding input dimension is used by the model—a higher value can therefore be interpreted as a rough measure of feature importance. For deeper layers, however, the hierarchical nature of neural networks means it is much harder to understand what a weight is applied to.

The analysis of model activations was used in multiple studies [93, 97, 120, 172, 186, 214, 227, 231]. This kind of inspection method usually involves visualizing the activations of the trained model over multiple examples, and thus inferring how different parts of the network react to known inputs. The input-perturbation network-prediction correlation map technique, introduced in [166], pushes this idea further by trying to identify causal relationships between the inputs and the decisions of a model. The impact of the perturbation on the activations of the last layer's units then shines light onto which characteristics of the input are important for the classifier to make a correct prediction. To do this, the input is first perturbed, either in the time- or frequency-domain, to alter its amplitude or phase characteristics [76], and then fed into the network. Occlusion sensitivity techniques [33, 102, 194] use a similar idea, by which the decisions of the network when different parts of the input are occluded are analyzed.

Several studies used backpropagation-based techniques to generate input maps that maximize activations of specific units [16, 158, 178, 207]. These maps can then be used to infer the role of specific neurons, or the kind of input they are sensitive to.

Finally, some model inspection techniques were used in a single study. For instance, in [53], the class activation map (CAM) technique was extended to overcome its limitations on

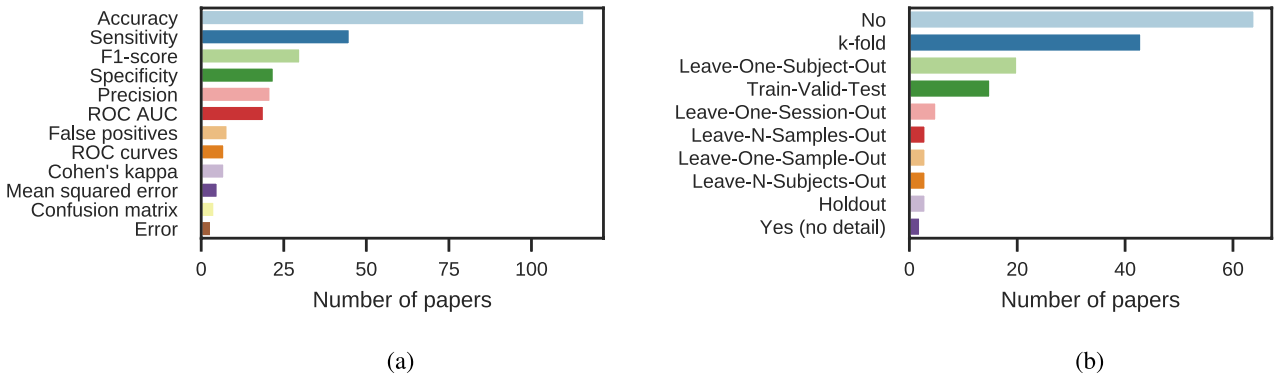

Figure 12. Performance metrics and cross-validation approaches. (a) Type of performance metrics used in the selected studies. Only metrics that appeared in at least three different studies are included in this figure. (b) Cross-validation approaches.

EEG data. To use CAMs in a CNN, the channel activations of the last convolutional layer must be averaged spatially before being fed into the model's penultimate layer, which is a FC layer. For a specific input image, a map can then be created to highlight parts of the image that contributed the most to the decision, by computing a weighted average of the last convolutional layer's channel activations. Other techniques include Deeplift [97], saliency maps [209], input-feature unit-output correlation maps [167], retrieval of closest examples [41], analysis of performance with transferred layers [71], analysis of most-activating input windows [76], analysis of generated outputs [75], and ablation of filters [97].

3.7. Reporting of results

The performance of DL methods on EEG is of great interest as it is still not clear whether DL can outperform traditional EEG processing pipelines [116]. Thus, a major question we thus aim to answer in this review is: 'Does DL lead to better performance than traditional methods on EEG?' However, answering this question is not straightforward, as benchmark datasets, baseline models, performance metrics and reporting methodology all vary considerably between the studies. In contrast, other application domains of DL, such as computer vision and NLP, benefit from standardized datasets and reporting methodology [61].

Therefore, to provide as satisfying an answer as possible, we adopt a two-pronged approach. First, we review how the studies reported their results by focusing on directly quantifiable items: (1) the type of baseline used as a comparison in each study, (2) the performance metrics, (3) the validation procedure, and (4) the use of statistical testing. Second, based on these points and focusing on studies that reported accuracy comparisons with baseline models, we analyze the reported performance of a majority of the reviewed studies.

3.7.1. Type of baseline. When contributing a new model, architecture or methodology to solve an already existing problem, it is necessary to compare the performance of the new model to the performance of state-of-the-art models commonly used for the problem of interest. Indeed, without a baseline comparison, it is not possible to assess whether the proposed method provides any advantage over the current state-of-the-art.

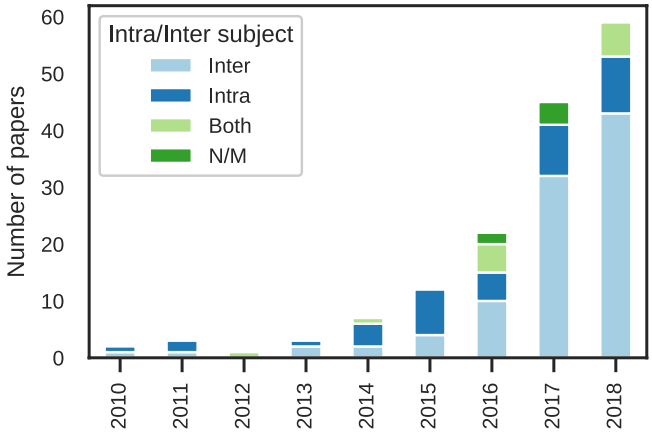

Figure 13. Distribution of intra- versus inter-subject studies per year.

Points of comparison are typically obtained in two different ways: (1) (re)implementing standard models or (2) referring to published models. In the first case, authors will implement their own baseline models, usually using simpler models, and evaluate their performance on the same task and in the same conditions. Such comparisons are informative, but often do not reflect the actual state-of-the-art on a specific task. In the second case, authors will instead cite previous literature that reported results on the same task and/or dataset. This second option is not always possible, especially when working on private datasets or tasks that have not been explored much in the past.

In the case of typical EEG classification tasks, state-of-the-art approaches usually involve traditional processing pipelines that include feature extraction and shallow/classical machine learning models. With that in mind, 68.2% of the studies selected included at least one traditional processing pipeline as a baseline model (see figure 15). Some studies instead (or also) compared their performance to DL-based approaches, to highlight incremental improvements obtained by using different architectures or training methodology: 34.4% of the studies therefore included at least one DL-based model as a baseline model. Out of the studies that did not compare their models to a baseline, six did not focus on the classification of EEG.

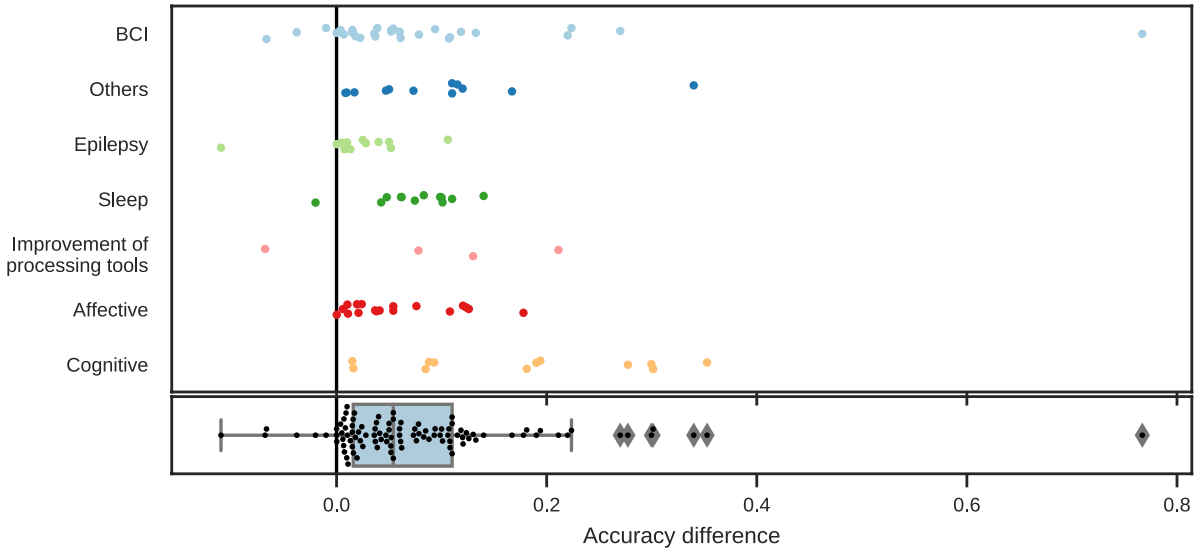

Figure 14. Difference in accuracy between each proposed DL model and corresponding baseline model for studies reporting accuracy (see section 3.7.6 for a description of the inclusion criteria). The difference in accuracy is defined as the difference between the best DL model and the best corresponding baseline. In the top figure, each study/task is represented by a single point, and studies are grouped according to their respective domains. The bottom figure is a box plot representing the overall distribution. *The result which achieved an accuracy difference of nearly* 77% *[179] was found to be caused by a flawed design in [107] and should therefore be considered as an outlier.*

Therefore, in total, 20.8% of the studies did not report baseline comparisons, making it impossible to assess the added value of their proposed methods in terms of performance.

3.7.2. Performance metrics. The types of performance metrics used by studies focusing on EEG classification are shown in figure 12(a). Unsurprisingly, most studies used metrics derived from confusion matrices, such as accuracy, sensitivity, f1-score, ROC AUC and precision. As highlighted in [33, 220], it is often preferable to use metrics that are robust to class imbalance, such as balanced accuracy, f1-score, and the ROC AUC for binary problems. This is often the case in sleep or epilepsy recordings, where clinical events are rare.

Studies that did not focus on the classification of EEG signals also mainly used accuracy as a metric. Indeed, these studies generally used a classification task to evaluate model performance, although their main purpose was different (e.g. correcting artifacts). In other cases, performance metrics specific to the study's purpose, such as generating data, were used, e.g. the inception score ([163]), the Fréchet inception distance ([83]), as well as custom metrics.

3.7.3. Validation procedure. When evaluating a machine learning model, it is important to measure its generalization performance, i.e. how well it performs on unseen data. In order to do this, it is common practice to divide the available data into a training and a test sets. When hyperparameters need to be tuned, the performance on the test set cannot be used anymore as an unbiased evaluation of the generalization performance of the model. Therefore, the training set is divided to obtain a third set called a 'validation set' which is used to select the best hyperparameter configuration, leaving the test set to evaluate the performance of the best model in an unbiased way. However, when the amount of data available is small, dividing the data into different sets and only using a subset for training can seriously undermine the performance of data-hungry models. A procedure known as 'cross-validation' is used in these cases, where the data is broken down into different partitions, which will then successively be used as either training or validation data.

The cross-validation techniques used in the selected studies are shown in figure 12(b). Some studies mentioned using cross-validation but did not provide any details. The category 'Train-Valid-Test' includes studies doing random permutations of train/valid, train/test or train/valid/test, as well as studies that mentioned splitting their data into training, validation and test sets but did not provide any details on the validation method. The Leave-One-Out variations correspond to the special case where $N = 1$ in the Leave-N-Out versions. 42% of the studies did not use any form of cross-validation. Interestingly, in [115], the authors proposed a 'warm restart' technique to improve performance and/or generalization of stochastic gradient descent and to relax the need to access a validation set by providing a recommendation solution as the latest solution of the latest completed cycle/restart.

3.7.4. Subject handling. Whether a study focuses on intra- or inter-subject classification has an impact on the performance. Intra-subject models, which are trained and used on the data of a single subject, often lead to higher performance since the model has less data variability to account for. However, this means the data the model is trained on is obtained from a single subject, and thus often comprises only a few recordings. In inter-subject studies, models generally see more data, as multiple subjects are included, but must contend with greater data variability, which introduces different challenges.

In the case of inter-subject classification, the choice of the validation procedure can have a big impact on the reported performance of a model. The Leave-N-Subject-Out procedure, which uses different subjects for training and for testing,

may lead to lower performance, but is applicable to real-life scenarios where a model must be used on a subject for whom no training data is available. In contrast, using k-fold cross-validation on the combined data from all the subjects often means that the same subjects are seen in both the training and testing sets. In the selected studies, 23 out of the 108 studies using an inter-subject approach used a Leave-N-Subjects-Out or Leave-One-Subjects-Out procedure.

In the selected studies, 26% focused only on intra-subject classification, 62% focused only on inter-subject classification, 8% focused on both, and 4% did not mention it. Obviously, 'N/M' studies necessarily fall under one of the three previous categories. The 'N/M' might be due to certain domains using a specific type of experiment (i.e. intra or inter-subject) almost exclusively, thereby obviating the need to mention it explicitly.

Figure 13 shows that there has been a clear trend over the last few years to leverage DL for inter-subject rather than intra-subject analysis. In [41], the authors used a large dataset and tested the performance of their model both on new (unseen) subjects and on known (seen) subjects. They obtained 38% accuracy on unseen subjects and 75% on seen subjects, showing that classifying EEG data from unseen subjects can be significantly more challenging than from seen ones.

In [203], the authors compared their model on both intra- and inter-subject tasks. Despite the former case providing the model with less less training data than the latter, it led to better results. In [70], the authors compared different DL models and showed that cross-subject (37 subjects) models always performed worse than within-subject models. In [141], a hybrid system trained on multiple subjects and then fine-tuned on subject-specific data led to the best performance. Finally, in [194], the authors compared their DNN to a state-of-the-art traditional approach and showed that deep networks generalize better, although their performance on intra-subject classification is still higher than on inter-subject classification.

3.7.5. Statistical testing. To assess whether a proposed model is actually better than a baseline model, it is useful to use statistical tests. In total, 19.5% of the selected studies used statistical tests to compare the performance of their models to baseline models. The tests most often used were Wilcoxon signed-rank tests, followed by ANOVAs.

3.7.6. Comparison of results. Although, as explained above, many factors make this kind of comparison imprecise, we show in this section how the proposed approaches and traditional baseline models compared, as reported by the selected studies.

We focus on a specific subset of the studies to make the comparison more meaningful. First, we focus on studies that report accuracy as a direct measure of task performance. As shown in figure 12(a), this includes the vast majority of the studies. Second, we only report studies which compared their models to a traditional baseline, as we are interested in whether DL leads to better results than non-DL approaches. This means studies which only compared their results to other DL approaches are not included in this comparison. Third, some studies evaluated their approach on more than one task or dataset. In this case, we report the results on the task that has the most associated baselines. If that is more than one, we either report all tasks, or aggregate them if they are very similar (e.g. binary classification of multiple mental tasks, where performance is reported for each possible pair of tasks). In the case of multimodal studies, we only report the performance on the EEG-only task, if it is available. Finally, when reporting accuracy differences, we focus on the difference between the best proposed model and the best baseline model, per task. Following these constraints, a total of 102 studies/tasks were left for our analysis.

Figure 14 shows the difference in accuracy between each proposed model and corresponding baseline per domain type (as categorized in figure 4), as well as the corresponding distribution over all included studies and tasks.

The median gain in accuracy with DL is of 5.4%, with an interquartile range of 9.4%. Only four values were negative values, meaning the proposed DL approach led to a lower performance than the baseline. We notice a slight, although not significant, difference in the median accuracy difference of the preprint and peer-reviewed groups (4.7% and 6.00%), respectively; Mann–Whitney test $p = 0.072$). While this difference is minor and exemplifies the same trend of slightly higher performance of DL models over traditional methods, it might originate from the lower publication standards of non-peer-reviewed research.

The highest improvement in accuracy (76.7%), obtained in [179], was shown to be caused by flawed experimental design and preprocessing strategy in a replication study [107]. Therefore, the improvement obtained in [228] (35.3% on a mental workload level classification task) was the highest achieved in the articles reviewed. In that study, a naive Bayes classifier trained on various features (including spectral and information theoretic features) preceded by a principal component analysis (PCA), was used as baseline.

3.8. Reproducibility

Reproducibility is a cornerstone of science [122]: having reproducible results is fundamental to moving a field forward, especially in a field like machine learning where new ideas spread very quickly. Here, we evaluate ease with which the results of the selected papers can be reproduced by the community using two key criteria: the availability of their data and the availability of their code.

From the 154 studies reviewed, 53% used public data, 42% used private data[13], and 4% used both public and private data. In particular, studies focusing on BCI, epilepsy, sleep and affective monitoring made use of openly available datasets the most (see table 6). Interestingly, in cognitive monitoring, no publicly available datasets were used, and papers in that field all relied on internal recordings.

[13] Data that is not freely available online was considered private regardless of when and where it was recorded. Moreover, three of the reviewed studies mentioned that their data was available upon request but were included in the 'private' category.

Table 6. Most often used datasets by domain. Datasets that were only used by one study are grouped under 'Other' for each category.

Main domain	Dataset	# articles	References
Affective	DEAP [91]	9	[6, 19, 49, 88, 106, 111, 113, 114, 220]
	SEED [246]	3	[114, 239, 247]
BCI	BCI Competition [26, 27, 160]	10	[32, 44, 50, 97, 120, 161, 162, 167, 189, 223]
	Other	6	[8, 71, 77, 97, 167, 185]
	eegmmidb [165]	8	[43, 67, 118, 132, 235, 241, 243, 245]
	Keirn & Aunon (1989)[a]	2	[139, 146]
Cognitive	Other	2	[69, 92]
	EEG Eye State[b]	1	[129]
Epilepsy	Bonn University [10]	7	[2, 4, 85, 128, 135, 190, 204]
	CHB-MIT [174]	7	[141, 194, 199, 200, 202, 203, 231]
	TUH [73]	5	[57–59, 171, 224]
	Other	3	[58, 192, 200]
	Freiburg Hospital[c]	2	[199, 200]
Generation of data	BCI Competition [26, 27, 160]	2	[40, 238]
	MAHNOB [177]	1	[212]
	Other	1	[170]
	SEED [247]	1	[212]
Improvement of processing tools	BCI Competition [26, 27, 160]	3	[183, 221, 222]
	Other	3	[133, 182, 225]
	Bonn University [10]	1	[215]
	CHB-MIT [174]	1	[215]
	MAHNOB [177]	1	[46]
Others	TUH [73]	3	[157, 166, 244]
	eegmmidb [165]	3	[240, 242, 244]
	Other	2	[206, 244]
	EEG Eye State[b]	1	[101]
Sleep	MASS [137]	4	[33, 45, 150, 186]
	Sleep EDF [89]	4	[186, 201, 209, 219]
	Other	3	[55, 178, 198]
	UCDDB[d]	3	[95, 96, 121]

[a] www.cs.colostate.edu/eeg/main/data/1989_Keirn_and_Aunon
[b] https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State
[c] http://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database
[d] https://physionet.org/pn3/ucddb/

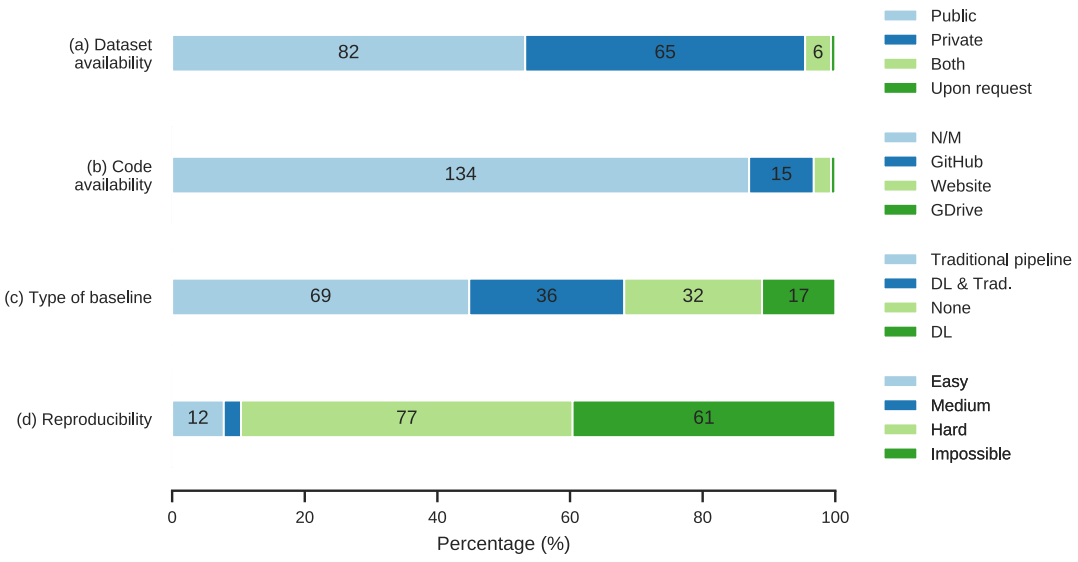

Figure 15. Reproductibility of the selected studies. (a) Availability of the datasets used in the studies, (b) availability of the code, shown by where the code is hosted, (c) type of baseline used to evaluate the performance of the trained models and (d) estimated reproducibility level of the studies (Easy: both the data and the code are available, Medium: the code is available but some data is not publicly available, Hard: either the code or the data is available but not both, Impossible: neither the data nor the code are available).

Fittingly, a total of 33 papers (21%) explicitly mentioned that more publicly available data is required to support research on DL-EEG. In clinical settings, the lack of labeled data, rather than the quantity of data, was specifically pointed out as an obstacle.

As for the source code, only 20 papers (13%) chose to make it available online [16, 92, 95, 97, 115, 166, 167, 170, 178, 179, 181, 182, 186, 214, 217, 240–244] and as illustrated in figure 15, GitHub is by far the preferred code sharing platform. Needless to say, having access to the source code behind published results can drastically reduce time and increase incentive to reproduce a paper's results.

Therefore, taking both data and code availability into account, only 12 out of 154 studies (8%) could easily be reproduced using both the same data and code [95, 115, 166, 170, 178, 179, 182, 186, 214, 240, 241, 243]. 4 out of 154 studies (3%) shared their code but tested on both private and public data making their studies only partially reproducible [97, 167, 242, 244], see figure 15. As follows, a significant number of studies (61) did not have publicly available data or code, making them almost impossible to reproduce.

It is important to note, moreover, that for the results of a study to be perfectly reproduced, the authors would also need to share the weights (i.e. parameters) of the network. Sharing the code and the architecture of the network might not be sufficient since retraining the network could converge to a different minimum. On the other hand, retraining the network could also end up producing better results if a better performing model is obtained. For recommendations on how to best share the results, the code, the data and relevant information to make a study easy to reproduce, please see the discussion section and the checklist provided in appendix B.

4. Discussion

In this section, we review the most important findings from our results section, and discuss the significance and impact of various trends highlighted above. We also provide recommendations for DL-EEG studies and present a checklist to ensure reproducibility in the field.

4.1. Rationale

It was expected that most papers selected for the review would focus on the classification of EEG data, as DL has historically led to important improvements on supervised classification problems [98]. Interestingly though, several papers also focused on new applications that were made possible or facilitated by DL: for instance, generating images conditioned on EEG, generating EEG, transfer learning between subjects, or feature learning. One of the main motivations for using DL cited by the papers reviewed was the ability to use raw EEG with no manual feature extraction steps. We expect these kinds of applications that go beyond using DL as a replacement for traditional processing pipelines to gain in popularity.

4.2. Data

A critical question concerning the use of DL with EEG data remains 'How much data is enough data?'. In section 3.3, we explored this question by looking at various descriptive dimensions: the number of subjects, the amount of EEG recorded, the number of training/test/validation examples, the sampling rate and data augmentation schemes used.

Although a definitive answer cannot be reached, the results of our meta-analysis show that the amount of data necessary to at least match the performance of traditional approaches is already available. Out of the 154 papers reviewed, only six reported lower performance for DL methods over traditional benchmarks. To achieve these results with limited amounts of data, shallower architectures were often preferred. Data augmentation techniques were also used successfully to improve performance when only limited data was available. However, more work is required to clearly assess their advantages and disadvantages. Indeed, although many studies used overlapping sliding windows, there seems to be no consensus on the best overlapping percentage to use, e.g. the impact of using a sliding window with 1% overlap versus 95% overlap is still not clear. BCI studies had the highest variability for this hyperparameter, while clinical applications such as sleep staging already appeared more standardized with most studies using 30 s non-overlapping windows.

Many authors concluded their paper suggesting that having access to more data would most likely improve the performance of their models. With large datasets becoming public, such as the TUH Dataset [73] and the National Sleep Research Resource [236], deeper architectures similar to the ones used in computer vision might become increasingly usable. However, it is important to note that the availability of data is quite different across domains. In clinical fields such as sleep and epilepsy, data usually comes from hospital databases containing years of recordings from several patients, while other fields usually rely on data coming from lab experiments with a limited number of subjects.

The potential of DL in EEG also lies in its ability (at least in theory) to generalize across subjects and to enable transfer learning across tasks and domains. Although intra-subject models still work best when only limited data is available, given the inherent subject variability of EEG data, transfer learning might be the key to moving past this limitation. Indeed, Page and colleagues [141] showed that with hybrid models, one can train a neural network on a pool of subjects and then fine-tune it on a specific subject, achieving good performances without needing as much data from a specific subject.

While the amount of data is critical in achieving high performance on machine learning tasks (and particularly for deep learning), the *quality* of the data is also very important. In many fields of application of DL, input data usually has a high SNR: in both CV and NLP, for instance, virtually noise-free images and natural language excerpts are easy to obtain. EEG data, on the other hand, can accumulate noise at many different levels, which makes learning from it much harder. Most often, once the data is recorded, the noise is impossible or

very difficult to mitigate. With that in mind, high quality and well-maintained hardware is crucial to collecting clean EEG data, however the capacity of the experimenter to prepare and use the equipment properly will ultimately determine signal quality. Prepping participants to ensure their compliance with the recording protocol is also fundamental to obtaining meaningful data. Similarly, reliable recording requires well planned out experimental design, including stimulus presentation when applicable. Furthermore, while naturally modulated by its end purpose, the quality of the data is influenced by its diversity, e.g. how many different individuals and how different they are. A balanced number of examples in each class can also drastically improve the usefulness of a large dataset. In brief, we believe both the quantity and the quality of the data must be taken into account when assessing the usefulness of a dataset, which is particularly true with electrophysiological data.

While we did report the sampling rate, we did not investigate its effect on performance because no relationship stood out particularly in any of the reviewed papers. The impact of the number of channels though, was specifically studied. For example, in [33], the authors showed that they could achieve comparable results with a lower number of channels. As shown in figure 8(a), a few studies used low-cost EEG devices, typically limited to a lower number of channels. These more accessible devices might therefore benefit from DL methods, but could also enable faster data collection on a larger-scale, thus facilitating DL in return.

As DL-EEG is highly data-driven, it is important when publishing results to clearly specify the amount of data used and to clarify terminology (see table 1 for an example). We noticed that many studies reviewed did not clearly describe the EEG data that they used (e.g. the number of subjects, number of sessions, window length to segment the EEG data, etc) and therefore made it hard or impossible for the reader to evaluate the work and compare it to others. Moreover, reporting learning curves (i.e. performance as a function of the number of examples) would give the reader valuable insights on the bias and variance of the model.

4.3. EEG processing

According to our findings, the great majority of the reviewed papers preprocessed the EEG data before feeding it to the deep neural network or extracting features. Despite observing this trend, we also noticed that recent studies outperformed their respective baseline(s) using completely raw EEG data. Almogbel *et al* [7] used raw EEG data to classify cognitive workload in vehicle drivers, and their best model achieved a classification accuracy approximately 4% better than their benchmarks which employed preprocessing on the EEG data. Similarly, Aznan *et al* [12] outperformed the baselines by a 4% margin on SSVEP decoding using no preprocessing. Thus, the answer to whether it is necessary to preprocess EEG data when using DNNs remains elusive.

As most of the works considered did not use, or explicitly mention using, artifact removal methods, it appears that this EEG processing pipeline step is in general not required. However, one should observe that in specific cases such as tasks that inherently elicit quick eye movements (MATB-II [38]), artifact handling might still be crucial to obtaining desired performance.

One important aspect we focused on is whether it is necessary to use EEG features as inputs to DNNs. After analyzing the type of input used by each paper, we observed that there was no clear preference for using features or raw EEG time-series as input. We noticed though that most of the papers using CNNs used raw EEG as input. With CNNs becoming increasingly popular, one can conclude that there is a trend towards using raw EEG instead of hand-engineered features. This is not surprising, as we observed that one of the main motivations mentioned for using DNNs on EEG processing is to automatically learn features. Furthermore, frequency-based features, which are widely used as hand-crafted features in EEG [116], are very similar to the temporal filters learned by a CNN. Indeed, these features are often extracted using Fourier filters which apply a convolutive operation. This is also the case for the temporal filters learned by a CNN although in the case of CNNs the filters are learned.

From our analysis, we also aimed to identify which input type should be used when trying to solve a problem from scratch. While the answer depends on many factors such as the domain of application, we observed that in some cases raw EEG as input consistently outperformed baselines based using classically extracted features. For example, for seizure classification, recently proposed models using raw EEG data as input [72, 138, 204] achieved better performances than classical baseline methods, such as SVMs with frequency-domain features. For this particular task, we believe following the current trend of using raw EEG data is the best way to start exploring a new approach.

4.4. Deep learning methodology

Another major topic this review aimed at covering is the DL methodology itself. Our analysis focused on architecture trends and training decisions, as well as on model selection techniques.

4.4.1. Architecture. Given the inherent temporal structure of EEG, we expected RNNs would be more widely employed than models that do not explicitly take time dependencies into account. However, almost half of the selected papers used CNNs. This observation is in line with recent discussions and findings regarding the effectiveness of CNNs for processing time series [13]. We also noticed that the use of energy-based models such as RBMs has been decreasing, whereas on the other hand, popular architectures in the computer vision community such as GANs have started to be applied to EEG data as well. As suggested by a Kruskal–Wallis test ($p = 0.043$), the choice of architecture seems to have had an impact on the reported accuracy improvement over traditional baselines: in the reviewed papers, CNNs and DBNs generally led to higher improvements, while AE-based models led to the lowest

improvements. Although this might reflect an actual advantage of convolutional architectures or of DBN-based unsupervised pretraining over vanilla recurrent or fully connected architectures, the considerable variability in the experiments reported in the reviewed papers makes it impossible to draw any conclusion yet. Instead, we believe focused studies will be necessary to evaluate the impact of architectural choices on performance on a domain-by-domain basis.

Moreover, regarding architecture depth, most of the papers used fewer than five layers. When comparing this number with popular object recognition models such as VGG and ResNet for the ImageNet challenge comprising 19 and 34 layers respectively, we conclude that for EEG data, shallower networks are currently necessary. Schirrmeister *et al* [196] specifically focused on this aspect, comparing the performance of architectures with different depths and structures, such as fully convolutional layers and residual blocks, on different tasks. Their results showed that in most cases, shallower fully convolutional models outperformed their deeper counterpart and architectures with residual connections. However, the authors later found the weight initialization to be critical in successfully training deeper architectures such as ResNet on an intracranial task [210], suggesting hyperparameter tuning might be key to using deeper architectures on neurophysiological data (personal communication, April 17, 2019).

4.4.2. Training and optimization. Although crucial to achieving good results when using neural networks, only 20% of the papers employed some hyperparameter search strategy. Even fewer studies provided detailed information about the method used. Amongst these, Stober *et al* [182] described their hyperparameter selection method and cited its corresponding implementation; in addition, the available budget in number of iterations per searching trial as well as the cross-validation split were mentioned in the paper.

4.4.3. Model inspection. Inspecting trained DL models is important, as DNNs are notoriously seen as black boxes, when compared to more traditional methods. Indeed, straightforward model inspection techniques such as visualizing the weights of a linear classifier are not applicable to deep neural networks; their decisions are thus much harder to understand. This is problematic in clinical settings for instance, where understanding and explaining the choice made by a classification model might be critical to making informed clinical choices. Neuroscientists might also be interested by what drives a model's decisions and use that information to shape hypotheses about brain function.

Although it can manifest with any machine model based on elaborate EEG features, the problem of identifying whether or not informative patterns stem from brain or artifactual activity is exacerbated by DL. Especially when considering end-to-end models trained on raw data (which is the case of almost half of the studies included in this review), any pattern correlated with the target of the learning task might end up being used by a model to drive decisions. When no artifact handling is done (at least 46% of the studies), it then becomes likely that artifactual components, which are typically much stronger in amplitude than actual EEG sources, are being used somehow by a DL model. In many applications where the unique concern is classification performance (e.g. BCI, sleep staging, seizure detection) and for subjects for whom artifacts are robust covariates of the measured condition, this might not be problematic. However, if the end goal requires the system to solely rely on brain activity (e.g. BCI for locked-in individuals who can not rely on residual muscle activity or in neuroscience-specific investigations), it is necessary to implement artifact handling procedures and, as far as possible, inspect the models trained. Since artifactual signatures are usually well-characterized, it should be possible to use methods like those mentioned in table 5 to assess whether brain activity or artifacts drive decisions in a DL model.

About 27% of the reviewed papers looked at interpreting their models. Interesting work on the topic, specifically tailored to EEG, was reviewed in [53, 76, 167]. Sustained efforts aimed at inspecting models and understanding the patterns they rely on to reach decisions are necessary to broaden the use of DL for EEG processing.

4.5. Reported results

Our meta-analysis focused on how studies compared classification accuracy between their models and traditional EEG processing pipelines on the same data. Although a great majority of studies reported improvements over traditional pipelines, this result has to be taken with a grain of salt. First, the difference in accuracy does not tell the whole story, as an improvement of 10%, for example, is typically more difficult to achieve from 80% to 90% than from 40% to 50%. More importantly though, very few articles reported negative improvements, which could be explained by a publication bias towards positive results.

The reported baseline comparisons were highly variable: some used simple models (e.g. combining straightforward spectral features and linear classifiers), others used more sophisticated pipelines (including multiple features and non-linear approaches), while a few reimplemented or cited state-of-the-art models that were published on the same dataset and/or task. Often, the description of baseline models is also too succinct to effectively assess whether the baselines are optimal for a given task: for instance, the performance on the training set can be used to assess whether the baseline models are in the overfitting or underfitting regime. Since the observed improvement will likely be higher when comparing to simple baselines than to state-of-the-art results, the values that we report might be biased positively. For instance, only two studies used Riemannian geometry-based processing pipelines as baseline models [12, 97], although these methods have set a new state-of-the-art in multiple EEG classification tasks [116].

Moreover, many different tasks and thus datasets were used. These datasets are often private, meaning there is very limited or no previous literature reporting results on them. On top of this, the lack of reproducibility standards can lead to low accountability: study results are not expected to be

replicated and can be inflated by non-standard practices such as omitting cross-validation.

Notwithstanding the limits of the improvement in accuracy as a performance metric (as described above), we ran a series of non-parametric statistical tests to assess whether any of the collected data items seem to covary with accuracy improvement. We used Mann–Whitney rank-sum tests for binary data items, Kruskal–Wallis analysis of variance for data items with more than two possible values, and Spearman's rank correlation for numerical data items, and considered their p-value. All p-values were found to be above a significance level of 0.05, except for the data item 'Architecture'. This result was discussed in section 4.4.1 above. As for the other data items, the inconclusiveness of the statistical tests most likely stem from highly variable and imprecise baseline comparisons across studies. Therefore, the impact of each of these data items remains better described by well-controlled domain-specific (and even dataset-specific) studies which might not be generalizable across domains. We tried to highlight studies that reported such interesting comparisons in both the Results and Discussion sections of this review.

Different approaches have been taken to solve the problem of heterogeneity of result reporting and benchmarking in the field of machine learning. For instance, OpenML [208] is an online platform that facilitates the sharing and running of experiments, as well as the benchmarking of models. As of November 2018, the platform already contained one EEG dataset and multiple submissions. The MOABB [87], a solution tailored to the field of brain–computer interfacing, is a software framework for ensuring the reproducibility of BCI experiments and providing public benchmarks for many BCI datasets. In [82], a similar approach, but for DL specifically, is proposed.

Additionally, a few EEG/MEG/ECoG classification online competitions have been organized in the last years, for instance the Physionet challenge [52] or various competitions on the Kaggle platform (see table 1 of [39]). These competitions informally act as benchmarks: they provide a standardized dataset with training and test splits, as well as a leaderboard listing the performance achieved by every competitor. These platforms can then be used to evaluate the state-of-the-art as they provide a publicly available comparison point for new proposed architectures. For instance, the IEEE NER 2015 Conference competition on error potential decoding could have been used as a benchmark for the studies reviewed that focused on this topic. Generally speaking, rigorous studies of the impact of different methodologies on specific datasets will be necessary to set up clear benchmarks that can be built upon (e.g. [156] for the TUH Seizure corpus).

Making use of these tools, or extending them to other EEG-specific tasks, appears to be one of the greatest challenges for the field of DL-EEG at the moment, and might be the key to more efficient and productive development of practical EEG applications. Whenever possible, authors should make sure to provide as much information as possible on the baseline models they have used, and explain how to replicate their results (see section 4.6).

4.6. Reproducibility

The significant use of public EEG datasets across the reviewed studies suggests that open data has greatly contributed to recent developments in DL-EEG. On the other hand, 42% of studies used data not publicly available—notably in domains such as cognitive monitoring. To move the field forward, it is thus important to create new benchmark datasets and share internal recordings. Moreover, the great majority of papers did not make their code available. Many papers reviewed are thus more difficult to reproduce: the data is not available, the code has not been shared, and the baseline models that were used to compare the performances of the models are either non-existent or not available.

Recent initiatives to promote best practices in data and code sharing would benefit the field of DL-EEG. FAIR neuroscience [216] and the Brain Imaging Data Structure (BIDS) [64] both provide guidelines and standards on how to acquire, organize and share data and code. BIDS extensions specific to EEG [149] and MEG [130] were also recently proposed. Moreover, open source software toolboxes are available to perform DL experiments on EEG. For example, the recent toolbox developed by Schirrmeister and colleagues, called BrainDecode [167], enables faster and easier development cycles by providing the basic functionality required for DL-EEG analysis while offering high level and easy to use functions to the user. The use of common software tools could facilitate reproducibility in the community. Beyond reproducibility, we believe simplifying access to data, making domain knowledge accessible and sharing code will enable more people to jump into the field of DL-EEG and contribute, transforming what has traditionally been a domain-specific problem into a more general problem that can be tackled with machine learning and DL methods.

To move forward in that direction, we are planning a follow-up to this literature review in the form of an online public portal listing in greater detail results of published research on multiple openly available datasets. We believe such a portal will be critical in the advancement of the DL-EEG literature.

4.7. Recommendations

To improve the quality and reproducibility of the work in the field of DL-EEG, we propose six guidelines in table 7. Moreover, appendix B presents a checklist of items that are critical to ensuring reproducibility and should be included in future studies. A similar checklist, but specifically targeting machine learning publications, has also recently been proposed[14].

4.7.1. Supplementary material. Along with the current paper, we make our data items table and related code available online at http://dl-eeg.com. We encourage interested readers to consult it in order to dive deeper into data items that are of specific interest to them—it should be straightforward to reproduce and extend the results and figures presented in this review

[14] www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

Table 7. Recommendations for future DL-EEG studies. See appendix B for a detailed list of items to include.

	Recommendation	Description
1	Clearly describe the architecture	Provide a table or figure clearly describing your model (e.g. see [33, 59, 167])
2	Clearly describe the data used	Make sure the number of subjects, the number of examples, the data augmentation scheme, etc are clearly described. Use unambiguous terminology or define the terms used (for an example, see table 1)
3	Use existing datasets	Whenever possible, compare model performance on public datasets
4	Include state-of-the-art baselines	If focusing on a research question that has already been studied with traditional machine learning, clarify the improvements brought by using DL
5	Share internal recordings	Whenever possible
6	Share reproducible code	Share code (including hyperparameter choices and model weights) that can easily be run on another computer, and potentially reused on new data

using the code provided. The data item table is intended to be updated frequently with new articles, therefore results will be brought up to date periodically.

Authors of DL-EEG papers not included in the review are invited to submit a summary of their article following the format of our data items table to our online code repository. We also invite authors whose papers are already included in the review to verify the accuracy of our summary. Eventually, we would like to indicate which studies have been submitted or verified by the original authors.

By updating the data items table regularly and inviting researchers in the community to contribute, we hope to keep the supplementary material of the review relevant and up-to-date as long as possible.

4.8. Limitations

In this section, we quickly highlight some limitations of the present work. First, although the search methodology used to identify relevant studies is well-founded, it undeniably did not capture all of the existing literature on the topic. Therefore, we have abstained from drawing absolute conclusions on the different data items, and instead focused on highlighting trends. As described in section 2, our search terms were not biased toward any type of architecture, and so we are confident the results we present in this review are sound.

Second, our decision to include preprints from arXiv and bioRxiv in the database search requires some justification. It is important to note that preprints are not peer-reviewed. Therefore, some of the studies we selected might not be of the same quality and scientific rigor as the ones coming from peer-reviewed journals or conferences. For this reason, whenever a preprint was followed by a publication in a peer-reviewed venue, we focused our analysis on the peer-reviewed version. Nonetheless, we did not find significant differences between the preprints and the peer-reviewed studies in terms of reported improvement in accuracy. ArXiv has been largely adopted by the DL community as a means to quickly disseminate results and encourage fast research iteration cycles. Since the field of DL-EEG is still young and a limited number of publications was available at the time of writing, we decided to include all the papers we could find, knowing that some of the newer trends would be mostly visible in repositories such as arXiv. Our goal with this review was to provide a transparent and objective analysis of the trends in DL-EEG. By including preprints, we feel we provided a better view of the current state-of-the-art, and are also in a better position to give recommendations on how to share results of DL-EEG studies moving forward.

Third, in order to keep this review reasonable in length, we decided to focus our analysis on the points that we judged most interesting and valuable. As a result, various factors that impact the performance of DL-EEG were not covered in the review. For example, we did not cover weight initialization: in [59], the authors compared 10 different initialization methods and showed an impact on the specificity metric, with ranged from 85.1% to 96.9%. Similarly, multiple data items were collected during the review process, but were not included in the analysis. These items, which include data normalization procedures, software toolboxes, hyperparameter values, loss functions, training hardware, training time, etc, remain available online for the interested reader. We are confident other reviews or research articles will be able to focus on more specific elements.

Finally, as any literature review in a field that is quickly evolving, the relevance of our analysis decays with time as new articles are being published and new trends are established. Since our last database search, we have already identified other articles that should eventually be added to the analysis. Again, making this work a living review by providing the data and code online will hopefully ensure the review will be of value and remain relevant for years to come.

5. Conclusion

The usefulness of EEG as a functional neuroimaging tool is unequivocal: clinical diagnosis of sleep disorders and epilepsy, monitoring of cognitive and affective states, as well as brain–computer interfacing all rely heavily on the analysis of EEG. However, various challenges remain to be solved. For instance, time-consuming tasks currently carried out by human experts, such as sleep staging, could be automated to increase the availability and flexibility of EEG-based diagnosis. Additionally, better generalization performance between subjects will be necessary to truly make BCIs useful. DL has been proposed as a potential candidate to tackle these

challenges. Consequently, the number of publications applying DL to EEG processing has seen an exponential increase over the last few years, clearly reflecting a growing interest in the community in these kinds of techniques.

In this review, we highlighted current trends in the field of DL-EEG by analyzing 154 studies published between January 2010 and July 2018 applying DL to EEG data. We focused on several key aspects of the studies, including their origin, rationale, the data they used, their EEG processing methodology, DL methodology, reported results and level of reproducibility.

Among the major trends that emerged from our analysis, we found that (1) DL was mainly used for classifying EEG in domains such as brain–computer interfacing, sleep, epilepsy, cognitive and affective monitoring, (2) the quantity of data used varied a lot, with datasets ranging from 1 to over 16 000 subjects (mean = 223; median = 13), producing 62 up to 9750 000 examples (mean = 251 532; median = 14 000) and from two to 4800 000 min of EEG recording (mean = 62 602; median = 360), (3) various architectures have been used successfully on EEG data, with CNNs, followed by RNNs and AEs, being most often used, (4) there is a clear growing interest towards using raw EEG as input as opposed to handcrafted features, (5) almost all studies reported a small improvement from using DL when compared to other baselines and benchmarks (median = 5.4%), and (6) while several studies used publicly available data, only a handful shared their code—the great majority of studies reviewed thus cannot easily be reproduced.

This review also shows that more targeted work needs to be done around the amount of data required to fully exploit the potential advantages of DL in EEG processing. Such work could explore the relationship between performance and the amount of data, the relationship between performance and data augmentation and the relationship between performance, the amount of data and the depth of the network.

Moreover, given the high variability in how results were reported, we made six recommendations to ensure reproducibility and fair comparison of results: (1) clearly describe the architecture, (2) clearly describe the data used, (3) use existing datasets, whenever possible, (4) include state-of-the-art baselines, ideally using the original authors' code, (5) share internal recordings, whenever possible, and (6) share code, as it is the best way to allow others to pick up where your work leaves off. We also provided a checklist (see appendix B) to help authors of DL-EEG studies make sure all the relevant information is available in their publications to allow straightforward reproduction.

Finally, to help the DL-EEG community maintain an up-to-date list of published work, we made our data items table open and available online. The code to reproduce the statistics and figures of this review as well as the full summaries of the papers are also available at http://dl-eeg.com.

The current general interest in artificial intelligence and DL has greatly benefited various fields of science and technology. Advancements in other field of application will most likely benefit the neuroscience and neuroimaging communities in the near future, and enable more pervasive and powerful applications based on EEG processing. We hope this review will constitute a good entry point for EEG researchers interested in applying DL to their data, as well as a good summary of the current state of the field for DL researchers looking to apply their knowledge to new types of data. A planned follow-up to this review will be an online portal providing clear and reproducible benchmarks for deep learning-based analysis of EEG data, accessible at http://dl-eeg.com.

Acknowledgments

We thank Raymundo Cassani, Colleen Gillon, João Monteiro and William Thong for comments that greatly improved the manuscript.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC-RDC) for JF and YR (reference number: RDPJ 514052-17), NSERC research funds for JF, HB, IA and THF, the Fonds québécois de la recherche sur la nature et les technologies (FRQNT) for YR and InteraXon Inc. (graduate funding support) for HB.

Appendix A. List of acronyms

AE	autoencoder.
BCI	brain–computer interface.
CCA	canonical correlation analysis.
CNN	convolutional neural network
CV	computer vision.
DBN	deep belief network.
DL	deep learning.
DNN	deep neural network.
ECoG	electrocorticography.
EEG	electroencephalography.
EMG	electromyography.
EOG	electroculography.
ERP	event-related potential.
FC	fully-connected.
GAN	generative adversarial network.
ICA	independent component analysis.
LSTM	long short-term memory.
MEG	magnetoencephalography.
NLP	natural language processing.
PCA	principal component analysis.
PSD	power spectral density.
RBM	restricted Boltzmann machine.
RNN	recurrent neural network.
ROC AUC	area under the receiver operating curve.
RSVP	rapid serial visual presentation.
SDAE	stacked denoising autoencoder.
SGD	stochastic gradient descent.
SNR	signal-to-noise ratio.
STFT	short-time Fourier transform.
SVM	support vector machine.
wICA	wavelet-enhanced independent component analysis.

Appendix B. Checklist of items to include in a DL-EEG study

This section contains a checklist of items we believe DL-EEG papers should mention to ensure their published results are readily reproducible. The following items of information should all be clearly stated at one point or another in the text or supplementary materials of future DL-EEG studies:

B.1. Data

- ☐ Number of subjects (and relevant demographic data)
- ☐ Electrode montage including reference(s) (number of channels and their locations)
- ☐ Shape of one example (e.g. '256 samples × 16 channels')
- ☐ Data augmentation technique (e.g. percentage of overlap for sliding windows)
- ☐ Number of examples in training, validation and test sets

B.2. EEG processing

- ☐ Temporal filtering, if any
- ☐ Spatial filtering, if any
- ☐ Artifact handling techniques, if any
- ☐ Resampling, if any

B.3. Neural network architecture

- ☐ Architecture type
- ☐ Number of layers (consider including a diagram or table to represent the architecture)
- ☐ Number of learnable parameters

B.4. Training hyperparameters

- ☐ Parameter initialization
- ☐ Loss function
- ☐ Batch size
- ☐ Number of epochs
- ☐ Stopping criterion
- ☐ Regularization (e.g. dropout, weight decay, etc)
- ☐ Optimization algorithm (e.g. stochastic gradient descent, Adam, RMSProp, etc)
- ☐ Learning rate schedule and optimizer parameters
- ☐ Values of all hyperparameters (including random seed) for the results that are presented in the paper
- ☐ Hyperparameter search method

B.5. Performance and model comparison

- ☐ Performance metrics (e.g. f1-score, accuracy, etc)
- ☐ Type of validation scheme (intra- versus inter-subject, leave-one-subject-out, k-fold cross-validation, etc)
- ☐ Description of baseline models (thorough description or reference to published work)

ORCID iDs

Yannick Roy https://orcid.org/0000-0003-4408-5221

References

[1] Aboalayon K A I, Faezipour M, Almuhammadi W S and Moslehpour S 2016 Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation *Entropy* **18** 272
[2] Acharya U R, Oh S L, Hagiwara Y, Tan J H and Adeli H 2017 Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals *Computers in Biology and Medicine* pp 1–9
[3] Acharya U R, Sree S V, Swapna G, Martis R J and Suri J S 2013 Automated EEG analysis of epilepsy: a review *Knowl.-Based Syst.* **45** 147–65
[4] Ahmedt-Aristizabal D, Fookes C, Nguyen K and Sridharan S 2018 Deep classification of epileptic signals (arXiv:1801.03610)
[5] Al-Nafjan A, Hosny M, Al-Ohali Y and Al-Wabil A 2017 Review and classification of emotion recognition based on EEG brain–computer interface system research: a systematic review *Appl. Sci.* **7** 1239
[6] Alhagry S, Fahmy A A and El-Khoribi R A 2017 Emotion recognition based on EEG using LSTM recurrent neural network *Int. J. Adv. Comput. Sci. Appl.* **8** 8–11
[7] Almogbel M A, Dang A H and Kameyama W 2018 EEG-signals based cognitive workload detection of vehicle driver using deep learning *20th Int. Conf. on Advanced Communication Technology* vol 7 pp 256–9
[8] An J and Cho S 2016 Hand motion identification of grasp-and-lift task from electroencephalography recordings using recurrent neural networks *Int. Conf. on Big Data and Smart Computing, BigComp 2016* pp 427–9
[9] An X, Kuang D, Guo X, Zhao Y and He L 2014 A deep learning method for classification of EEG data based on motor imagery *Lecture Notes Comput. Sci.* **8590 LNBI** 203–10 (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)
[10] Andrzejak R G, Lehnertz K, Mormann F, Rieke C, David P and Elger C E 2001 Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state *Phys. Rev.* E **64** 061907
[11] Arns M, Conners C K and Kraemer H C 2013 A decade of EEG theta/beta ratio research in ADHD: a meta-analysis *J. Attention Disorders* **17** 374–83
[12] Aznan N K N, Bonner S, Connolly J D, Moubayed N A and Breckon T P 2018 On the classification of SSVEP-based dry-EEG signals via convolutional neural networks (arXiv:1805.04157)
[13] Bai S, Kolter J Z and Koltun V 2018 An empirical evaluation of generic convolutional and recurrent networks for sequence modeling (arXiv:1803.01271)
[14] Baltatzis V, Bintsi K M, Apostolidis G K and Hadjileontiadis L J 2017 Bullying incidences identification within an immersive environment using HD EEG-based analysis: a swarm decomposition and deep learning approach *Sci. Rep.* **7** 17292
[15] Bashivan P, Rish I and Heisig S 2016 Mental state recognition via wearable EEG (arXiv:1602.00985)
[16] Bashivan P, Rish I, Yeasin M and Codella N 2015 Learning representations from EEG with deep recurrent-convolutional neural networks (arXiv:1511.06448)
[17] Behncke J, Schirrmeister R T, Burgard W and Ball T 2017 The signature of robot action success in EEG signals of a

human observer: decoding and visualization using deep convolutional neural networks (arXiv:1711.06068)
[18] Ben-David S, Blitzer J, Crammer K and Pereira F 2007 Analysis of representations for domain adaptation *Advances in Neural Information Processing Systems* pp 137–44
[19] Ben Said A, Mohamed A, Elfouly T, Harras K and Wang Z J 2017 Multimodal deep learning approach for Joint EEG-EMG data compression and classification *IEEE Wireless Communications and Networking Conf.*
[20] Bergstra J and Bengio Y 2012 Random search for hyper-parameter optimization *J. Mach. Learn. Res.* **13** 281–305
[21] Berka C, Levendowski D J, Lumicao M N, Yau A, Davis G, Zivkovic V T, Olmstead R E, Tremoulet P D and Craven P L 2007 {EEG} correlates of task engagement and mental workload in vigilance, learning, and memory tasks *Aviat. Space Environ. Med.* **78** B231–44
[22] Biasiucci A, Franceschiello B and Murray M M 2019 Electroencephalography *Curr. Biol.* **29** R80–5
[23] Bigdely-Shamlo N, Mullen T, Kothe C, Su K M and Robbins K A 2015 The PREP pipeline: standardized preprocessing for large-scale EEG analysis *Frontiers Neuroinform.* **9** 16
[24] Bishop C M 1995 *Neural Networks for Pattern Recognition* vol 92 (Oxford: Oxford University Press)
[25] Biswal S, Kulas J, Sun H, Goparaju B, Westover M B, Bianchi M T and Sun J 2017 SLEEPNET: automated sleep staging system via deep learning 1–17 (arXiv:1707.08262)
[26] Blankertz B *et al* 2004 The bci competition 2003 *IEEE Trans. Biomed. Eng.* **51** 1044–51
[27] Blankertz B, Muller K R, Krusienski D J, Schalk G, Wolpaw J R, Schlogl A, Pfurtscheller G, Millan J R, Schroder M and Birbaumer N 2006 The bci competition iii: validating alternative approaches to actual bci problems *IEEE Trans. Neural Syst. Rehabil. Eng.* **14** 153–9
[28] Brock A, Donahue J and Simonyan K 2018 Large scale gan training for high fidelity natural image synthesis (arXiv:1809.11096)
[29] Bu N, Shima K and Tsuji T 2010 EEG discrimination using wavelet packet transform and a reduced-dimensional recurrent neural network *Proc. of the 10th IEEE Int. Conf. on Information Technology and Applications in Biomedicine* pp 1–4
[30] Castellanos N P and Makarov V A 2006 Recovering EEG brain signals: artifact suppression with wavelet enhanced independent component analysis *J. Neurosci. Methods* **158** 300–12
[31] Cecotti H and Gräser A 2011 Convolutional neural networks for P300 detection with application to brain–computer interfaces *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 433–45
[32] Cecotti H, Eckstein M P and Giesbrecht B 2014 Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering *IEEE Trans. Neural Netw. Learn. Syst.* **25** 2030–42
[33] Chambon S, Galtier M N, Arnal P J, Wainrib G, Gramfort A, Paristech T and Nov M L 2017 A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series *IEEE Trans. Neural Syst. Rehabil. Eng.* **26** 758–69
[34] Chiarelli A M, Croce P, Merla A and Zappasodi F 2018 Deep learning for hybrid EEG-fNIRS brain–computer interface: application to motor imagery classification *J. Neural Eng.* **15** 036028
[35] Chu L, Qiu R, Liu H, Ling Z, Zhang T and Wang J 2017 Individual recognition in schizophrenia using deep learning methods with random forest and voting classifiers: insights from resting state EEG streams (arXiv:1707.03467)
[36] Clerc M, Bougrain L and Lotte F 2016 *Brain–Computer Interfaces 1: Foundations and Methods* (New York: Wiley)
[37] Cole S R and Voytek B 2018 Cycle-by-cycle analysis of neural oscillations (*bioRxiv*: 302000)
[38] Comstock J R 1994 Mat-multi-attribute task battery for human operator workload and strategic behavior research
[39] Congedo M, Barachant A and Bhatia R 2017 Riemannian geometry for EEG-based brain–computer interfaces; a primer and a review *Brain–Comput. Interfaces* **4** 155–74
[40] Corley I A and Huang Y 2018 Deep EEG super-resolution: upsampling EEG spatial resolution with generative adversarial networks *IEEE EMBS Int. Conf. on Biomedical & Health Informatics* pp 4–7
[41] Deiss O, Biswal S, Jin J, Sun H, Westover M B and Sun J 2018 HAMLET: interpretable human and machine co-learning technique (arXiv:1803.09702)
[42] Russakovsky O *et al* 2015 ImageNet large scale visual recognition challenge *Int. J. Comput. Vis.* **115** 211–52
[43] Dharamsi T, Das P, Pedapati T, Bramble G, Muthusamy V, Samulowitz H, Varshney K R, Rajamanickam Y, Thomas J and Dauwels J 2017 Neurology-as-a-service for the developing world (arXiv:1711.06195)
[44] Ding S, Zhang N, Xu X, Guo L and Zhang J 2015 Deep extreme learning machine and its application in EEG classification *Math. Problems Eng.* **2015** 129021
[45] Dong H, Supratak A, Pan W, Wu C, Matthews P M and Guo Y 2018 Mixed neural network approach for temporal sleep stage classification *IEEE Trans. Neural Syst. Rehabil. Eng.* **26** 324–33
[46] Drouin-Picaro A and Falk T H 2016 Using deep neural networks for natural saccade classification from electroencephalograms *Proc. IEEE EMBS Int. Student Conf.: Expanding the Boundaries of Biomedical Engineering and Healthcare* (IEEE) pp 1–4
[47] Duchi J, Hazan E and Singer Y 2011 Adaptive subgradient methods for online learning and stochastic optimization *J. Mach. Learn. Res.* **12** 2121–59
[48] Engemann D A *et al* 2018 Robust EEG-based cross-site and cross-protocol classification of states of consciousness *Brain* **141** 3179–92
[49] Frydenlund A and Rudzicz F 2015 Emotional affect estimation using video and EEG data in deep neural networks *Lecture Notes Comput. Sci.* **9091** 273–80 (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)
[50] Gao G, Shang L, Xiong K, Fang J, Zhang C and Gu X 2018 EEG classification based on sparse representation and deep learning *NeuroQuantology* **16** 789–95
[51] Gao Y, Lee H J and Mehmood R M 2015 Deep learninig of EEG signals for emotion recognition *IEEE Int. Conf. on Multimedia and Expo Workshops* (IEEE) pp 1–5
[52] Ghassemi M M, Moody B E, Lehman L, Song C, Li Q, Sun H, Mark R G, Westover M B and Clifford G D 2018 You snooze, you win: the physionet/computing in cardiology challenge 2018 *Comput. Cardiol.* **45** (http://www.cinc.org/archives/2018/pdf/CinC2018-049.pdf)
[53] Ghosh A, Dal Maso F, Roig M, Mitsis G D and Boudrias M H 2018 Deep semantic architecture with discriminative feature visualization for neuroimage analysis (arXiv:1805.11704)
[54] Giacino J T, Fins J J, Laureys S and Schiff N D 2014 Disorders of consciousness after acquired brain injury: the state of the science *Nat. Rev. Neurol.* **10** 99
[55] Giri E P, Fanany M I and Arymurthy A M 2016 Combining generative and discriminative neural networks for sleep stages classification (arXiv:1610.01741)
[56] Giri E P, Fanany M I and Arymurthy A M 2016 Ischemic stroke identification based on EEG and EOG using 1D convolutional neural network and batch normalization (arXiv:1610.01757)
[57] Golmohammadi M, Torbati A H H N, de Diego S L, Obeid I and Picone J 2017 Automatic analysis of EEGs

using big data and hybrid deep learning architectures (arXiv:1712.09771)
[58] Golmohammadi M, Ziyabari S, Shah V, de Diego S L, Obeid I and Picone J 2017 Deep architectures for automated seizure detection in scalp EEGs (arXiv:1712.09776)
[59] Golmohammadi M, Ziyabari S, Shah V, Von Weltin E, Campbell C, Obeid I and Picone J 2017 Gated recurrent networks for seizure detection *IEEE Signal Processing in Medicine and Biology Symp.* pp 1–5
[60] Goodfellow I 2016 Nips 2016 tutorial: generative adversarial networks (arXiv:1701.00160)
[61] Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* vol 1 (Cambridge, MA: MIT Press)
[62] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Advances in Neural Information Processing Systems* pp 2672–80
[63] Gordienko Y, Stirenko S, Kochura Y, Alienin O, Novotarskiy M and Gordienko N 2017 Deep learning for fatigue estimation on the basis of multimodal human–machine interactions (arXiv:1801.06048)
[64] Gorgolewski K J *et al* 2016 The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments *Sci. Data* **3** 160044
[65] Gramfort A, Strohmeier D, Haueisen J, Hämäläinen M S and Kowalski M 2013 Time-frequency mixed-norm estimates: sparse M/EEG imaging with non-stationary source activations *NeuroImage* **70** 410–22
[66] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A 2017 Improved training of wasserstein GANs *Advances in Neural Information Processing Systems* pp 5767–77
[67] Alomari M H, Samaha A and AlKamha K 2013 Automated classification of L/R hand movement EEG signals using advanced feature extraction and machine learning *Int. J. Adv. Comput. Sci. Appl.* **4** 6
[68] Hagihira S 2015 Changes in the electroencephalogram during anaesthesia and their physiological basis *Br. J. Anaesthesia* **115** i27–31
[69] Hajinoroozi M, Mao Z and Huang Y 2015 Prediction of driver's drowsy and alert states from EEG signals with deep learning *IEEE 6th Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing* pp 493–6
[70] Hajinoroozi M, Mao Z and Lin Y P 2017 Deep transfer learning for cross-subject and cross-experiment prediction of image rapid serial visual presentation events from EEG data *Int. Conf. on Augmented Cognition* vol 10284 pp 45–55
[71] Hajinoroozi M, Mao Z, Jung T P, Lin C T and Huang Y 2016 EEG-based prediction of driver's cognitive performance by deep convolutional neural network *Signal Process.: Image Commun.* **47** 549–55
[72] Hao Y, Khoo H M, von Ellenrieder N, Zazubovits N and Gotman J 2018 DeepIED: an epileptic discharge detector for EEG-fMRI based on deep learning *NeuroImage* **17** 962–75
[73] Harati A, López S, Obeid I and Picone J 2014 The TUH EEG CORPUS: a big data resource for automated EEG interpretation *IEEE Signal Processing in Medicine and Biology Symp.* (IEEE) pp 1–5
[74] Hari R and Puce A 2017 *MEG-EEG Primer* (Oxford: Oxford University Press)
[75] Hartmann K G, Schirrmeister R T and Ball T 2018 EEG-GAN: generative adversarial networks for electroencephalograhic (EEG) brain signals (arXiv:1806.01875)
[76] Hartmann K G, Schirrmeister R T and Ball T 2018 Hierarchical internal representation of spectral features in deep convolutional networks trained for EEG decoding *6th Int. Conf. on Brain–Computer Interface* (IEEE) pp 1–6
[77] Hasib M M, Nayak T and Huang Y 2018 A hierarchical LSTM model with attention for modeling EEG non-stationarity for human decision prediction *IEEE EMBS Int. Conf. on Biomedical and Health Informatics* pp 104–7
[78] He B, Sohrabpour A, Brown E and Liu Z 2018 Electrophysiological source imaging: a noninvasive window to brain dynamics *Annu. Rev. Biomed. Eng.* **20** 171–96
[79] He K, Zhang X, Ren S and Sun J 2015 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 770–8
[80] Hefron R, Borghetti B, Schubert Kabban C, Christensen J and Estepp J 2018 Cross-participant EEG-based assessment of cognitive workload using multi-path convolutional recurrent neural networks *Sensors* **18** 1339
[81] Hefron R G, Borghetti B J, Christensen J C and Kabban C M S 2017 Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation *Pattern Recognit. Lett.* **94** 96–104
[82] Heilmeyer F A, Schirrmeister R T, Fiederer L D J, Völker M, Behncke J and Ball T 2018 A large-scale evaluation framework for EEG deep learning architectures (arXiv:1806.07741)
[83] Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S 2017 GANs trained by a two time-scale update rule converge to a local nash equilibrium *Advances in Neural Information Processing Systems* pp 6626–37
[84] Hohman F M, Kahng M, Pienta R and Chau D H 2018 Visual analytics in deep learning: an interrogative survey for the next frontiers *IEEE Trans. Vis. Comput. Graph.* pp 2674–93
[85] Hussein R, Palangi H, Ward R and Wang Z J 2018 Epileptic seizure detection: a deep learning approach (arXiv:1803.09848)
[86] Jas M, Engemann D A, Bekhti Y, Raimondo F and Gramfort A 2017 Autoreject: automated artifact rejection for MEG and EEG data *NeuroImage* **159** 417–29
[87] Jayaram V and Barachant A 2018 MOABB: trustworthy algorithm benchmarking for BCIs *J. Neural Eng.* **15** 066011
[88] Jirayucharoensak S, Pan-Ngum S and Israsena P 2014 EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation *Sci. World J.* **2014** 627892
[89] Kemp B, Zwinderman A H, Tuk B, Kamphuisen H A and Oberye J J 2000 Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG *IEEE Trans. Biomed. Eng.* **47** 1185–94
[90] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)
[91] Koelstra S, Mühl C, Soleymani M, Lee J S, Yazdani A, Ebrahimi T, Pun T, Nijholt A and Patras I 2012 DEAP: a database for emotion analysis; using physiological signals *IEEE Trans. Affective Comput.* **3** 18–31
[92] Kuanar S, Athitsos V, Pradhan N, Mishra A and Rao K R 2018 Cognitive analysis of working memory load from EEG, by a deep recurrent neural network *IEEE Signal Process. Soc.* pp 2576–80
[93] Kwak N S, Müller K R and Lee S W 2017 A convolutional neural network for steady state visual evoked potential classification under ambulatory environment *PLoS One* **12** 1–20
[94] Kwon Y, Nan Y and Kim S D 2017 Transformation of EEG signal for emotion analysis and dataset construction for DNN learning *Lecture Notes in Electrical Engineering* vol 474 (Berlin: Springer) pp 96–101
[95] Längkvist M and Loutfi A 2018 A deep learning approach with an attention mechanism for automatic sleep stage classification (arXiv: 1805.05036)
[96] Längkvist M, Karlsson L and Loutfi A 2012 Sleep stage classification using unsupervised feature learning *Adv. Artif. Neural Syst.* **2012** 107046

[97] Lawhern V J, Solon A J and Waytowich N R 2018 EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces *J. Neural Eng.* **15** 056013

[98] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436

[99] LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W and Jackel L D 1989 Backpropagation applied to handwritten zip code recognition *Neural Comput.* **1** 541–51

[100] LeCun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proc. IEEE* **86** 2278–324

[101] Lee W H, Ortiz J, Ko B and Lee R 2018 Time series segmentation through automatic feature learning (arXiv:1801.05394)

[102] Lee Y and Huang Y 2018 Generating target/non-target images of an RSVP experiment from brain signals in by conditional generative adversarial network *2018 IEEE EMBS International Conference on Biomedical & Health Informatics* (BHI) pp 182–5

[103] Li F, Zhang G, Wang W, Xu R, Schnell T, Wen J, McKenzie F and Li J 2017 Deep models for engagement assessment with scarce label information *IEEE Trans. Hum.-Mach. Syst.* **47** 598–605

[104] Li J and Cichocki A 2014 Deep learning of multifractal attributes from motor imagery induced EEG *Neural Information Processing* pp 503–10

[105] Li J, Struzik Z, Zhang L and Cichocki A 2015 Feature learning from incomplete EEG with denoising autoencoder *Neurocomputing* **165** 23–31

[106] Li K, Li X, Zhang Y and Zhang A 2013 Affective state recognition from EEG with deep belief networks *Proc.—2013 IEEE Int. Conf. on Bioinformatics and Biomedicine* pp 305–10

[107] Li R, Johansen J S, Ahmed H, Ilyevsky T V, Wilbur R B, Bharadwaj H M and Siskind J M 2018 Training on the test set? An analysis of Spampinato *et al* (arXiv:1609.00344)

[108] Li X, Grandvalet Y and Davoine F 2018 Explicit inductive bias for transfer learning with convolutional networks (arXiv:1802.01483)

[109] Li X, Zhang P, Song D, Yu G, Hou Y and Hu B 2015 EEG based emotion identification using unsupervised deep feature learning *SIGIR2015 Workshop on Neuro-Physiological Methods in IR Research* pp 2–4

[110] Li Y, Schwing A, Wang K C and Zemel R 2017 Dualing GANs *Advances in Neural Information Processing Systems* pp 5606–16

[111] Li Z, Tian X, Shu L, Xu X and Hu B 2018 Emotion recognition from EEG Using RASM and LSTM *Internet Multimedia Computing and Service* vol 819 (Berlin: Springer)

[112] Liao C Y, Chen R C and Tai S K 2018 Emotion stress detection using EEG signal and deep learning technologies *IEEE Int. Conf. on Applied System Invention* pp 90–3

[113] Lin W, Li C and Sun S 2017 Deep convolutional neural network for emotion recognition using EEG and peripheral physiological signal *Int. Conf. on Image and Graphics* pp 385–94

[114] Liu W, Zheng W L and Lu B L 2016 Multimodal emotion recognition using multimodal deep learning

[115] Loshchilov I and Hutter F 2016 SGDR: stochastic gradient descent with warm restarts (arXiv:1608.03983)

[116] Lotte F, Bougrain L and Clerc M 2015 *Electroencephalography (EEG)-Based Brain–Computer Interfaces* (American Cancer Society) pp 1–20

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

[118] Major T C and Conrad J M 2017 The effects of pre-filtering and individualizing components for electroencephalography neural network classification *Conf. Proc.—IEEE SOUTHEASTCON*

[119] Makeig S, Bell A J, Jung T P and Sejnowski T J 1996 Independent component analysis of electroencephalographic data *Advances in Neural Information Processing Systems* vol **8** pp 145–51

[120] Manor R and Geva A B 2015 Convolutional neural network for multi-category rapid serial visual presentation BCI *Frontiers Comput. Neurosci.* **9** 1–12

[121] Manzano M, Guillén A, Rojas I and Herrera L J 2017 Deep learning using EEG data in time and frequency domains for sleep stage classification *Lecture Notes Comput. Sci.* **10305 LNCS** 132–41 (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)

[122] Munafò M R, Nosek B A, Bishop D V M, Button K S, Chambers C D, du Sert N P, Simonsohn U, Wagenmakers E-J, Ware J J and Ioannidis J P A 2017 A manifesto for reproducible science *Nat. Hum. Behav.* **1** 0021

[123] Mehmood R M, Du R and Lee H J 2017 Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors *IEEE Access* **5** 14797–806

[124] Mohamed A K, Marwala T and John L R 2011 Single-trial EEG discrimination between wrist and finger movement imagery and execution in a sensorimotor BCI *Ann. Int. Conf. Proc. IEEE Eng. Med. Biol. Soc.* pp 6289–93

[125] Moinnereau M A, Brienne T, Brodeur S, Rouat J, Whittingstall K and Plourde E 2018 Classification of auditory stimuli from EEG signals with a regulated recurrent neural network reservoir

[126] Morabito F C, Campolo M, Ieracitano C, Ebadi J M, Bonanno L, Bramanti A, Desalvo S, Mammone N and Bramanti P 2016 Deep convolutional neural networks for classification of mild cognitive impaired and Alzheimer's disease patients from scalp EEG recordings *IEEE 2nd Int. Forum on Research and Technologies for Society and Industry Leveraging a Better Tomorrow*

[127] Morabito F C *et al* 2017 Deep learning representation from electroencephalography of early-stage Creutzfeldt–Jakob disease and features for differentiation from rapidly progressive dementia *Int. J. Neural Syst.* **27** 1650039

[128] Naderi M A and Mahdavi-Nasab H 2010 Analysis and classification of EEG signals using spectral analysis and recurrent neural networks *17th Iranian Conf. of Biomedical Engineering* (IEEE) pp 1–4

[129] Narejo S, Pasero E and Kulsoom F 2016 EEG based eye state classification using deep belief network and stacked autoencoder *Int. J. Electr. Comput. Eng.* **6** 3131–41

[130] Niso G *et al* 2018 Meg-bids, the brain imaging data structure extended to magnetoencephalography *Sci. Data* **5** 180110

[131] Nolan H, Whelan R and Reilly R B 2010 FASTER: fully automated statistical thresholding for EEG artifact rejection *J. Neurosci. Methods* **192** 152–62

[132] Normand R and Ferreira H A 2015 Superchords: the atoms of thought (arXiv:150501228)

[133] Nurse E, Mashford B S, Yepes A J, Kiral-Kornek I, Harrer S and Freestone D R 2016 Decoding EEG and LFP signals using deep learning: heading TrueNorth *Proc. of the ACM Int. Conf. on Computing Frontiers* pp 259–66

[134] O 'shea A, Lightbody G, Boylan G and Temko A 2017 Neonatal seizure detection using convolutional neural networks (arXiv:1709.05849)

[135] Omerhodzic I, Avdakovic S, Nuhanovic A and Dizdarevic K 2013 Energy distribution of EEG signals: EEG signal wavelet-neural network classifier 2 (arXiv:1307.7897)

[136] Oord A V D *et al* 2017 Parallel wavenet: fast high-fidelity speech synthesis (arXiv:1711.10433)
[137] O'reilly C, Gosselin N, Carrier J and Nielsen T 2014 Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research *J. Sleep Res.* **23** 628–35
[138] O'Shea A, Lightbody G, Boylan G and Temko A 2018 Investigating the impact of CNN depth on neonatal seizure detection performance (arXiv: 1806.03044)
[139] Padmanabh L, Shastri R and Biradar S 2017 Mental tasks classification using EEG signal, discrete wavelet transform and neural network *Discovery* **48** 38–41
[140] Paez A 2017 Gray literature: an important resource in systematic reviews *J. Evidence-Based Med.* **10** 233–40
[141] Page A, Shea C and Mohsenin T 2016 Wearable seizure detection using convolutional neural networks with transfer learning *IEEE Int. Symp. on Circuits and Systems* pp 1086–9
[142] Palazzo S, Spampinato C, Kavasidis I, Giordano D and Shah M 2017 Generative adversarial networks conditioned by brain signals *Proc. IEEE Int. Conf. on Computer Vision* pp 3430–8
[143] Pardede J, Turnip M, Manalu D R and Turnip A 2015 Adaptive recurrent neural network for reduction of noise and estimation of source from recorded EEG signals *ARPN J. Eng. Appl. Sci.* **10** 993–7
[144] Parekh V, Subramanian R, Roy D and Jawahar C V 2018 An EEG-based image annotation system *Commun. Comput. Inf. Sci.* **841** 303–13
[145] Patanaik A, Ong J L, Gooley J J, Ancoli-Israel S and Chee M W L 2018 An end-to-end framework for real-time automatic sleep stage classification *Sleep* **41** 1–11
[146] Patnaik S, Moharkar L and Chaudhari A 2017 Deep RNN learning for EEG based functional brain state inference *Int. Conf. on Advances in Computing, Communication and Control*
[147] Perez-Benitez J L, Perez-Benitez J A and Espina-Hernandez J H 2018 Development of a brain computer interface interface using multi-frequency visual stimulation and deep neural networks *Int. Conf. on Electronics, Communications and Computers* pp 18–24
[148] Perez L and Wang J 2017 The effectiveness of data augmentation in image classification using deep learning (arXiv:1712.04621)
[149] Pernet C R, Appelhoff S, Flandin G, Phillips C, Delorme A and Oostenveld R 2019 Bids-EEG: an extension to the brain imaging data structure (bids) specification for electroencephalography *Sci. Data* **6** 103
[150] Phan H, Andreotti F, Cooray N, Chen O Y and De Vos M 2019 Joint classification and prediction CNN framework for automatic sleep stage classification *IEEE Trans. Biomed. Eng.* **66** 1285–96
[151] Prechelt L 1998 Automatic early stopping using cross validation: quantifying the criteria *Neural Netw.* **11** 761–7
[152] Radford A, Wu J, Child R, Luan D, Amodei D and Sutskever I 2018 Language models are unsupervised multitask learners *Technical Report* OpenAi
[153] Raposo F, de Matos D M, Ribeiro R, Tang S and Yu Y 2017 Towards deep modeling of music semantics using EEG regularizers (arXiv:1712.05197)
[154] Robbins H and Monro S 1951 A stochastic approximation method *Statistics* (Berlin: Springer) pp 102–9
[155] Rosenblatt F 1958 The perceptron: a probabilistic model for information storage and organization *Psychol. Rev.* **65** 386–408
[156] Roy S, Asif U, Tang J and Harrer S 2019 Machine learning for seizure type classification: setting the benchmark
[157] Roy S, Kiral-Kornek I and Harrer S 2018 ChronoNet: a deep recurrent neural network for abnormal EEG identification (arXiv:1802.00308)
[158] Ruffini G, Ibanez D, Castellano M, Dubreuil L, Gagnon J F, Montplaisir J and Soria-Frisch A 2018 Deep learning with EEG spectrograms in rapid eye movement behavior disorder (*bioRxiv*: 240267)
[159] Rumelhart D E, Hinton G E and Williams R J 1986 Learning representations by back-propagating errors *Nature* **323** 533
[160] Sajda P, Gerson A, Muller K R, Blankertz B and Parra L 2003 A data analysis competition to evaluate machine learning algorithms for use in brain-computer interfaces *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 184–5
[161] Sakhavi S and Guan C 2017 Convolutional neural network-based transfer learning and knowledge distillation using multi-subject data in motor imagery BCI *Int. IEEE/EMBS Conf. on Neural Engineering* pp 588–91
[162] Sakhavi S, Guan C and Yan S 2015 Parallel convolutional-linear neural network for motor imagery classification *23rd European Signal Processing Conf.* pp 2736–40
[163] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A and Chen X 2016 Improved techniques for training GANs *Advances in Neural Information Processing Systems* pp 2234–42
[164] Saon G *et al* 2017 English conversational telephone speech recognition by humans and machines (arXiv:1703.02136)
[165] Schalk G, McFarland D J, Hinterberger T, Birbaumer N and Wolpaw J R 2004 Bci2000: a general-purpose brain-computer interface (bci) system *IEEE Trans. Biomed. Eng.* **51** 1034–43
[166] Schirrmeister R T, Gemein L, Eggensperger K, Hutter F and Ball T 2017 Deep learning with convolutional neural networks for decoding and visualization of EEG pathology (arXiv:1708.08012)
[167] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420
[168] Schmidhuber J 2015 Deep learning in neural networks: an overview: read section 6.6 *Neural Netw.* **61** 85–117
[169] Schomer D L and Da Silva F L 2012 *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields* (Baltimore, MD: Williams & Wilkins)
[170] Schwabedal J T C, Snyder J C, Cakmak A, Nemati S and Clifford G D 2018 Addressing class imbalance in classification problems of noisy signals by using fourier transform surrogates (arXiv:1806.08675)
[171] Shah V, Golmohammadi M, Ziyabari S, Von Weltin E, Obeid I and Picone J 2017 Optimizing channel selection for seizure detection *Signal Processing in Medicine and Biology Symp.* (IEEE) pp 1–5
[172] Shamwell J, Lee H, Kwon H, Marathe A R, Lawhern V and Nothwang W 2016 Single-trial EEG RSVP classification using convolutional neural networks *Micro-and Nanotechnology Sensors, Systems, and Applications VIII* vol **9836**
[173] Shang J, Yuanyue H, Haixiang G, Yijing L, Mingyun G and Gong B 2017 Learning from class-imbalanced data: Review of methods and applications *Expert Syst. Appl.* **73** 220–39
[174] Shoeb A H and Guttag J V 2010 Application of machine learning to epileptic seizure detection *Proc. of the 27th Int. Conf. on Machine Learning* pp 975–82
[175] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition (arXiv:1409.1556)

[176] Snoek J, Larochelle H and Adams R P 2012 Practical bayesian optimization of machine learning algorithms *Advances in Neural Information Processing Systems* pp 2951–9
[177] Soleymani M, Lichtenauer J, Pun T and Pantic M 2012 A multimodal database for affect recognition and implicit tagging *IEEE Trans. Affective Comput.* **3** 42–55
[178] Sors A, Bonnet S, Mirek S, Vercueil L and Payen J F 2018 A convolutional neural network for sleep stage scoring from raw single-channel EEG *Biomed. Signal Process. Control* **42** 107–14
[179] Spampinato C, Palazzo S, Kavasidis I, Giordano D, Souly N and Shah M 2017 Deep learning human mind for automated visual classification *Proc.—30th IEEE Conf. on Computer Vision and Pattern Recognition* pp 4503–11
[180] Sree R A and Kavitha A 2017 Vowel classification from imagined speech using sub-band EEG frequencies and deep belief networks *4th Int. Conf. on Signal Processing, Communication and Networking* (IEEE) pp 1–4
[181] Stober S, Cameron D J and Grahn J A 2014 Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings *Neural Information Processing Systems* pp 1–9
[182] Stober S, Sternin A, Owen A M and Grahn J A 2015 Deep feature learning for EEG recordings (arXiv:1511.04306)
[183] Sturm I, Bach S, Samek W and Müller K R 2016 Interpretable deep neural networks for single-trial EEG classification (arXiv:1604.08201)
[184] Sun B and Saenko K 2016 Deep coral: correlation alignment for deep domain adaptation *European Conf. on Computer Vision* (Berlin: Springer) pp 443–50
[185] Sun P and Qin J 2016 Neural networks based EEG-speech models 1–10 (arXiv:1612.05369)
[186] Supratak A, Dong H, Wu C and Guo Y 2017 DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG *IEEE Trans. Neural Syst. Rehabil. Eng.* **25** 1998–2008
[187] Sutskever I, Hinton G, Krizhevsky A and Salakhutdinov R R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58
[188] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z 2016 Rethinking the inception architecture for computer vision *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 2818–26
[189] Tabar Y R and Halici U 2016 A novel deep learning approach for classification of EEG motor imagery signals *J. Neural Eng.* **14** 16003
[190] Talathi S S 2017 Deep recurrent neural networks for seizure detection and early seizure detection systems (arXiv:1706.03283)
[191] Tang Z, Li C and Sun S 2017 Single-trial EEG classification of motor imagery using deep convolutional neural networks *Optik* **130** 11–8
[192] Taqi A M, Al-Azzo F, Mariofanna M and Al-Saadi J M 2017 Classification and discrimination of focal and non-focal EEG signals based on deep neural network *Int. Conf. on Current Research in Computer Science and Information Technology* pp 86–92
[193] Teo J, Hou C L and Mountstephens J 2018 Preference classification using electroencephalography (EEG) and deep learning *J. Telecommun. Electron. Comput. Eng.* **10** 87–91
[194] Thodoroff P, Pineau J and Lim A 2016 Learning robust features using deep learning for automatic seizure detection (arXiv:1608.00220)
[195] Thorsten O Z and Christian K 2011 Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general *J. Neural Eng.* **8** 25005
[196] Tibor S R, Tobias S J, Josef F L D, Martin G, Katharina E, Michael T, Frank H, Wolfram B and Tonio B 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420
[197] Tieleman T, Hinton G E, Srivastava N and Swersky K 2012 Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude *COURSERA: Neural Netw. Mach. Learn.* **4** 26–31
[198] Tripathy R K and Rajendra Acharya U 2018 Use of features from RR-time series and EEG signals for automated classification of sleep stages in deep neural network framework *Biocybern. Biomed. Eng.* **38** 890–902
[199] Truong N D, Kuhlmann L, Bonyadi M R and Kavehei O 2018 Semi-supervised seizure prediction with generative adversarial networks (arXiv:1806.08235)
[200] Truong N D, Nguyen A D, Kuhlmann L, Bonyadi M R, Yang J, Ippolito S and Kavehei O 2018 Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram *Neural Netw.* **105** 104–11
[201] Tsinalis O, Matthews P M, Guo Y and Zafeiriou S 2016 Automatic sleep stage scoring with single-channel EEG using convolutional neural networks (arXiv:1610.01683)
[202] Tsiouris K M, Pezoulas V C, Zervakis M, Konitsiotis S, Koutsouris D D and Fotiadis D I 2018 A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals *Comput. Biol. Med.* **99** 24–37
[203] Turner J T, Page A, Mohsenin T and Oates T 2014 Deep belief networks used on high resolution multichannel electroencephalography data for seizure detection *AAAI Spring Symp. Series* pp 75–81
[204] Ullah I, Hussain M, Qazi E U H and Aboalsamh H 2018 An automated system for epilepsy detection using EEG brain signals based on deep learning approach (arXiv:1801.05412)
[205] Urigen J A and Garcia-Zapirain B 2015 EEG artifact removal state-of-the-art and guidelines *J. Neural Eng.* **12** 031001
[206] Van Putten M J A M, Olbrich S and Arns M 2018 Predicting sex from brain rhythms with deep learning *Sci. Rep.* 8
[207] van Putten M J, de Carvalho R and Tjepkema-Cloostermans M C 2018 Deep learning for detection of epileptiform discharges from scalp EEG recordings *Clin. Neurophysiol.* **129** e98–9
[208] Vanschoren J, van Rijn J N, Bischl B and Torgo L 2014 OpenML: networked science in machine learning *SIGKDD Explorations* **15** 49–60
[209] Vilamala A, Madsen K H and Hansen L K 2017 Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring (arXiv:1710.00633)
[210] Volker M, Hammer J, Schirrmeister R T, Behncke J, Fiederer L D, Schulze-Bonhage A, Marusic P, Burgard W and Ball T 2018 Intracranial error detection via deep learning *IEEE Int. Conf. on Systems, Man, and Cybernetics* (IEEE) pp 568–75
[211] Völker M, Schirrmeister R T, Fiederer L D J, Burgard W and Ball T 2017 Deep transfer learning for error decoding from non-invasive EEG *2018 6th Int. Conf. on Brain–Computer Interface (BCI)* (IEEE) pp 1–6
[212] Wang F, Zhong S H, Peng J, Jiang J and Liu Y 2018 Data augmentation for eeg-based emotion recognition with deep convolutional neural networks *Lecture Notes Comput. Sci.* **10705 LNCS** 82–93 (including subseries Lecture

Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)
[213] Wang S, Guo B, Zhang C, Bai X and Wang Z 2018 EEG detection and de-noising based on convolution neural network and Hilbert-Huang transform *Proc.—2017 10th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics* pp 1–6
[214] Waytowich N R, Lawhern V, Garcia J O, Cummings J, Faller J, Sajda P and Vettel J M 2018 Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials (arXiv:1803.04566)
[215] Wen T and Zhang Z 2018 Deep convolution neural network and autoencoders-based unsupervised feature learning of EEG signals *IEEE Access* **6** 25399–410
[216] Wilkinson M D 2016 Comment: The fair guiding principles for scientific data management and stewardship *Sci. Data* **3** 160018
[217] Wu Z, Wang H, Cao M, Chen Y and Xing E P 2018 Fair deep learning prediction for healthcare applications with confounder filtering 1–17
[218] Wulsin D F, Gupta J R, Mani R, Blanco J A and Litt B 2011 Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement *J. Neural Eng.* **8** 036015
[219] Xie S, Li Y, Xie X, Wang W and Duan X 2017 The analysis and classify of sleep stage using deep learning network from single-channel EEG signal *Lecture Notes Comput. Sci.* 10637 LNCS 752–8 (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)
[220] Xu H and Plataniotis K N 2016 Affective states classification using EEG and semi-supervised deep learning approaches *IEEE 18th Int. Workshop on Multimedia Signal Processing* pp 1–6
[221] Yang B, Duan K and Zhang T 2016 Removal of EOG artifacts from EEG using a cascade of sparse autoencoder and recursive least squares adaptive filter *Neurocomputing* **214** 1053–60
[222] Yang B, Duan K, Fan C, Hu C and Wang J 2018 Automatic ocular artifacts removal in EEG using deep learning *Biomed. Signal Process. Control* **43** 148–58
[223] Yang H, Sakhavi S, Ang K K and Guan C 2015 On the use of convolutional neural networks and augmented CSP features for multi-class motor imagery of EEG signals classification *Ann. Int. Conf. Proc. IEEE Eng. Med. Biol. Soc.* pp 2620–3
[224] Yang S, Golmohammadi M, Obeid I and Picone J 2016 Semi-automated annotation of signal events in clinical EEG data, Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA *Signal Processing in Medicine and Biology Symp.* pp 1–5
[225] Yepes A J, Tang J and Mashford B S 2017 Improving classification accuracy of feedforward neural networks for spiking neuromorphic chips *Int. Joint Conf. on Artificial Intelligence* pp 1973–9
[226] Yin Z and Zhang J 2016 Recognition of cognitive task load levels using single channel EEG and stacked denoising autoencoder *Chinese Control Conf.* (IEEE) pp 3907–12
[227] Yin Z and Zhang J 2017 Cross-session classification of mental workload levels using EEG and an adaptive deep learning model *Biomed. Signal Process. Control* **33** 30–47
[228] Yin Z and Zhang J 2017 Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights *Neurocomputing* **260** 349–66
[229] Yogatama D, Dyer C, Ling W and Blunsom P 2017 Generative and discriminative text classification with recurrent neural networks (arXiv:1703.01898)
[230] Yoon J, Lee J and Whang M 2018 Spatial and time domain feature of ERP speller system extracted via convolutional neural network *Comput. Intell. Neurosci.* **2018** 6058065
[231] Yuan Y, Xun G, Ma F, Suo Q, Xue H, Jia K and Zhang A 2018 A novel channel-aware attention framework for multi-channel EEG seizure detection via multi-view deep learning *IEEE EMBS Int. Conf. on Biomedical &* Health Informatics pp 4–7
[232] Zafar R, Dass S C and Malik A S 2017 Electroencephalogram-based decoding cognitive states using convolutional neural network and likelihood ratio based score fusion *PLoS ONE* **12** e0178410
[233] Zeiler M D 2012 ADADELTA: an adaptive learning rate method (arXiv:1212.5701)
[234] Zhang C, Bengio S, Hardt M, Recht B and Vinyals O 2016 Understanding deep learning requires rethinking generalization (arXiv:1611.03530)
[235] Zhang D, Yao L, Zhang X, Wang S, Chen W and Boots R 2018 Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface *32nd AAAI Conf. on Artificial Intelligence* pp 1703–10
[236] Zhang G Q, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D and Redline S 2018 The national sleep research resource: towards a sleep data commons. *J. Am. Med. Inform. Assoc.* **25** 1351–8
[237] Zhang J, Li S and Wang R 2017 Pattern recognition of momentary mental workload based on multi-channel electrophysiological data and ensemble convolutional neural networks *Frontiers Neurosci.* **11** 1–16
[238] Zhang Q and Liu Y 2018 Improving brain computer interface performance by data augmentation with conditional deep convolutional generative adversarial networks (arXiv:1806.07108)
[239] Zhang T, Zheng W, Cui Z, Zong Y and Li Y 2018 Spatial-temporal recurrent neural network for emotion recognition *IEEE Trans. Cybern.* **1**
[240] Zhang X, Yao L, Chen K, Wang X, Sheng Q and Gu T 2017 DeepKey: an EEG and gait based dual-authentication system (arXiv:1706.01606)
[241] Zhang X, Yao L, Huang C, Sheng Q Z and Wang X 2017 Intent recognition in smart living through deep recurrent neural networks *International Conference on Neural Information Processing* (Cham: Springer) pp 748–58
[242] Zhang X, Yao L, Kanhere S S, Liu Y, Gu T and Chen K 2018 Mindid: Person identification from brain waves through attention-based recurrent neural network *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* vol 2 pp 149:1–23
[243] Zhang X, Yao L, Sheng Q Z, Kanhere S S, Gu T and Zhang D 2018 Converting your thoughts to texts: enabling brain typing via deep feature learning of eeg signals *IEEE Int. Conf. on Pervasive Computing and Communications (PerCom)* pp 1–10
[244] Zhang X, Yao L, Wang X, Zhang W, Zhang S and Liu Y 2018 Know your mind: adaptive brain signal classification with reinforced attentive convolutional neural networks (arXiv:1802.03996)
[245] Zhang X, Yao L, Zhang D, Wang X, Sheng Q Z and Gu T 2017 Multi-person brain activity recognition via comprehensive EEG signal analysis *Proc. of the 14th EAI Int. Conf. on Mobile and Ubiquitous Systems: Computing, Networking and Services* (New York: ACM) pp 28–37

[246] Zheng W L and Lu B L 2015 Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks *IEEE Trans. Auton. Mental Dev.* **7** 162–75

[247] Zheng W L, Liu W, Lu Y, Lu B L and Cichocki A 2019 Emotionmeter: a multimodal framework for recognizing human emotions *IEEE Trans. Cybern.* **49** 1110–22

[248] Zheng W L, Zhu J Y, Peng Y and Lu B L 2014 EEG-based emotion classification using deep belief networks *Proc.—IEEE Int. Conf. on Multimedia and Expo* pp 1–6

[249] Zhou J and Xu W 2015 End-to-end learning of semantic role labeling using recurrent neural networks *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing (Volume 1: Long Papers) vol* 1 pp 1127–37