Deep learning applied to electroencephalogram data in mental disorders: A systematic review

Mateo de Bardeci a, b, c, Cheng Teng Ip d, e, Sebastian Olbrich a, b, c, *

ABSTRACT

In recent medical research, tremendous progress has been made in the application of deep learning (DL) techniques. This article systematically reviews how DL techniques have been applied to electroencephalogram (EEG) data for diagnostic and predictive purposes in conducting research on mental disorders. EEG-studies on psychiatric diseases based on the ICD-10 or DSM-V classification that used either convolutional neural networks (CNNs) or long-short-term-memory (LSTMs) networks for classification were searched and examined for the quality of the information they contained in three domains: clinical, EEG-data processing, and deep learning. Although we found that the description of EEG acquisition and pre-processing was sufficient in most of the studies, we found that many of them lacked a systematic characterization of clinical features. Furthermore, many studies used misguided model selection procedures or flawed testing. It is recommended that the study of psychiatric disorders using DL in the future must improve the quality of clinical data and follow state of the art model selection and testing procedures so as to achieve a higher research standard and head toward a clinical significance.

1. Introduction

For several decades, it has been claimed that progress in psychiatric research has been on the verge of creating a major transformation in the way mental disorders are diagnosed, treated and managed. Even as early as 1957, Eisenberg predicted that progress had been “exciting and substantial” and, referring to neurophysiological and sociocultural findings, “promise[d] […] gains in neuropharmacology.” This optimism was still being echoed 60 years later when the WPA-Lancet Psychiatry Commission wrote that “Psychiatry in the first quarter of the 21st century is at the cusp of major changes” and listed similar general advances (Bhugra et al., 2017). It is undeniable that past decades have seen significant progress in several neurobiological domains. Driven by the development of sophisticated imaging capabilities like MRI (Lerch et al., 2017), SPECT, and PET (Zipsky et al., 2007) as well as major breakthroughs in genetic research (Smoller et al., 2019) and new neuroscience techniques such as optogenetics (Tye & Deisseroth, 2012), our understanding of the underlying mechanisms and causes for various mental disorders has dramatically improved. At the clinical level, however, the way psychiatric practitioners diagnose mental disorders and decide what treatment might be best for patients has not kept pace with these technical advances (Ghaemi, 2018; North & Suris, 2017). While our understanding of neuropsychiatric disorders at the neuronal or even molecular level has increased, most clinically relevant decisions are still not based on objectifiable facts. Instead, these assessments are derived from mainly anamnestic features and experiences of the physician or therapist.

In light of this divergence, two separate important developments have taken place within recent years. First, it has been proposed to shift the focus from mainly diagnosis-based decisions for treatment toward a more individualized approach (Olbrich & Conradi, 2016). This process is similar to the precision medicine approach used in other fields of medicine (National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease, 2011). Driven by the use of the Research Criteria Domains (RDoC) approach announced in 2013 (Cutbert & Insel, 2013), much more attention has been focused on the...
value of using biological patterns to define common underlying causes and open possible treatment approaches. Second, there have been recently significant technological advances, especially a revival in the use of computational neural networks. The development of Deep Learning (DL), in particular the use of Convolutional Neural Network (CNNs) (Hinton et al., 2006) and Recursive Neural Networks (RNNs) as Long Short Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) in combination with the wider availability of affordable computational power has created a revolution in many scientific areas. The DL revolution has demonstrated an unprecedented level of accuracy in classifying, predicting, and natural language processing (NLP). While we have all witnessed the pervasive impact of these technological advances on our daily lives (as, e.g., in the availability of facial recognition on smartphones), these advances have enabled dramatic innovations in the medical sector: for example, the detection of retinal disease (Fauw et al., 2018) and the diagnostic classification of electrocardiogram data (Rathhal et al., 2016). Pattern recognition is one of the physician’s main tasks in the diagnostic process. The use of CNNs and LSTMs for pattern recognition has been shown to perform as good or even better than humans.

The combination of the two above described developments, leads to a shift in medicine towards more objective, data-driven and pattern-based diagnostic and therapeutic approaches. The research which has accompanied these developments, however, has been uneven and inconsistent. In the light of often announced “transformations in psychiatry” within the past decades, it is important at this stage to carefully and critically review how these emerging innovations are being applied to ensure that they conform to the highest scientific standards.

Besides imaging studies and genetic research, electrophysiology, especially since the first description of the electroencephalogram (EEG) (Berger, 1929), has always been widely used for the scientific analysis of neuropsychiatric disorders (Boutros, 2015, Hughes, 1996; Pogarell, 2017). While EEG has been widely adopted in pharaco-EEG research (e.g. the description of the key-lock principle by Itil (1983)), it is hardly ever applied in psychiatric routine clinical scenarios except for drug monitoring purposes or for ruling out the organic causes of syndromes. Although recent developments in the field of prognostic biomarkers seem promising (Olbrich et al., 2015; Rajpurkar et al., 2020; Rolle et al., 2020; Wu et al., 2020), a recent meta-analysis has raised serious concerns regarding their replicability (Widge et al., 2019). Still, DL based algorithms might push the usage of EEG in psychiatry by supporting the “Gestaltanalyse”, (which is the human ability to extract meaningful information for diagnostic, prognostic or even therapeutic usage from the highly complex EEG-timeseries) to a more automated, objective, and reliable level. The use of neural networks to perform these EEG classification tasks may lead a more efficient and effective way to identify previously unrecognized patterns in EEG traces and reveal their meaning while contributing to improve management of the underlying psychiatric syndromes.

In response to the developments described above, the present review aims to identify and evaluate the quality and the state of the current research being conducted which uses DL methods to analyse EEG time series for the study of mental disorders. To avoid the historical pitfall of overestimating the diagnostic or predictive power of these tools, the review specifically aims at determining the flaws and gaps in the way they are being applied. At the end of this review, a series of recommendations for improving the quality of DL usage to support diagnostic and prognostic tasks in psychiatry will be presented.

To achieve the goals, the current state of research in published papers has been systematically reviewed, evaluated, and categorized based on three domains:

1 Clinical Domain: The Clinical Domain includes all the relevant information to the clinical context, especially diagnostics entities and their underlying classification systems as reflected in the Diagnostic and Statistical Manual of Mental Disorders (DSM) or the International Classification of Disease (ICD). It also includes information regarding the description of diagnosis validation, the number of included subjects, and the population from which the sample was extracted, as well as any information on ongoing treatments, including medication and the type of clinical usage the paper is intended to address (i.e. a diagnostic approach, disease progress prediction or response prediction).

2 EEG-Data Domain: We include in the EEG-Data Domain the most important information on the recording of electrophysiological input data for the networks is included. This includes the conditions under which an EEG was recorded, which typically can be divided into resting state recordings, sleep recordings and recordings where the participants perform cognitive tasks. More technical variables include the sampling rate of the recording (i.e., the number of signal values stored per second) and the number of EEG channels that were used (the number of recording electrodes on the scalp). Other variables include the handling of the artefacts (technical and physiological signals removed that did not emerge from the brain) and the preprocessing in means of segmentation filtering and signal transformation.

3 Deep-Learning Domain: The Deep-Learning Domain includes information about feature engineering, network structure, optimisation algorithm, model selection, testing and backwards modelling. Feature engineering refers to the extraction of specific features from the data using domain knowledge in order to improve machine learning performance. Features of EEG data can typically be separated into four clusters (see Fig. 2): 1. Time domain; 2. Frequency domain (derived from Fourier or wavelet transformations); 3. Spatial features (2-dimensional activation maps derived from topographical maps); and 4. Connectivity measure matrices that report on the associations between different electrode sites. Further information of the DL domain includes the network structure and number of hidden layers.

Artificial Neural Networks (ANNs) are structured in layers with an input layer and an output layer, layers in between are called “hidden layers”. Further reported variables include the activation function (the mathematical operation that is performed between layers) and optimisation algorithm (being stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (Adam) the most popular optimization algorithms). Furthermore, the model selection and testing techniques are reported. To estimate the efficacy of the DL model, usually the model is tested in an independent dataset, or similarly, the initial data is split into independent train and test data. The test data should be used only once and should be kept “unseen” until testing. Therefore, to find the right network structure and to tune hyperparameters, the training data is usually split again into training and validation sets. By testing different network structures and hyperparameters in the validation set, the right model for testing can be determined and chosen. Finding the right model is known as “model selection”. Another common practice is to estimate efficacy is to use Cross Validation (CV). In CV the data is split into k independent folds and subsequently trained in k-1 folds and tested in the complementary fold until the model has been trained and tested in every fold (i.e., in the complete dataset). The case of k equals the number of datapoints is called Leave One Out Cross Validation (LOOCV). Similar to the independent test set approach, to keep the folds independent and “unseen”, CV can be performed only once. Otherwise, the performance may be overestimated. In many cases, a combination of CV and train-test split is used, where CV is used for model selection, and the best model is then tested in the independent test set. To evaluate the predictive power of a model, different scores can be used, accuracy being the most common one. Accuracy is defined as the proportion of correct predictions (both true positives and true negatives), but it neglects the proportion of false predictions. Therefore, when dealing with unbalanced data, the usage of other scores that account for the proportion of false predictions is encouraged. One alternative to accuracy is the F1-score, defined as the harmonic mean between precision and recall.
Finally, the Deep-Learning Domain includes backwards modelling, which is an approach to identify the features that were extracted in the process of training by the ANN. Performing backwards modelling can give important biological insights, as well as improve the interpretability of the results.

2. Methods

2.1. Inclusion criteria and inclusion procedure

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) procedure was used to complete the following steps (Moher et al., 2009). First, the PubMed database was queried for relevant papers written in English that had been published until 21st of October 2020. The search phrases used included terms for 1) electrophysiological measure (EEG), 2) DL networks, and 3) a variable term for mental disorders according to the F-code diagnosis of the International Classification of Diseases, version 10 (ICD-10, 1993, p. 10). This search yielded the following search term: "EEG AND (deep learning OR CNN OR convolutional net OR deep net OR deep belief OR LSTM OR deep boltzmann) AND xxx (keyword for mental disorder)". The variable keyword for the mental disorders represented the categories from F0 to F9 (see Table 1). From these papers, the references they used were in turn searched to identify further references and those additional references were searched in additional databases such as Google Scholar and the preprint server arXiv. All the papers identified from the first step were then checked for duplicates. When necessary, the full text was used to verify eligibility. At this point, only papers which used EEG, a clinical population, and a deep learning approach were retained for further reference.

All these papers were then assigned to one author (SO) to extract a list of variables from the three different domains (clinical, EEG-processing, DL-domains)). After this screening, all the papers deemed relevant were reviewed by a second author (CI or MB) who counter-checked the list of extracted variables. More papers had to be discarded at this point because they did not contain information or descriptions of the main features of the three domains for a classification (missing diagnosis, missing description of the EEG data, missing description of the used networks) (Fig. 1).

2.2. Exclusion criteria

In this review, papers referencing patients suffering from Parkinson’s disease or dementia were excluded from the list since they are coded as neurological conditions in ICD-10. Studies on the methodological improvement of sleep assessment using polysomnography and DL techniques were also excluded. Only those studies on diagnostic or prognostic questions regarding sleep disorders were considered.

2.3. Reproducibility

In the next step, we ranked the reproducibility of the DL approach of a study as poor, middle, or good. A paper was categorized as poor if at least one essential aspect was missing, and the DL would be impossible to reproduce; middle if all the essential aspects were described and reproducibility would be possible but still unspecific; and good if otherwise. Specifically, if any of the following characteristic of the DL process was not present, then that paper’s reproducibility was categorized as poor: input features, input dimensions, network structure and network type, number of hidden layers, activation functions, model selection technique and testing technique.

If the paper was not classified as poor, and if any of the following characteristic of the DL process was not present, then the paper’s reproducibility was categorized as middle: loss function and optimization algorithm.

If the paper contained all the required listed elements identified above, then we categorized the paper’s reproducibility as good.

2.4. Reporting in cases where different approaches and models were present

If a paper used multiple testing techniques, preprocessing procedures, and/or machine learning models, we gave priority to only one combination based on the following criteria:

1. The DL model over other machine learning models
2. The validity of its testing technique if present
3. The model which yielded the highest degree of accuracy

3. Results

From the corpus of 139 studies identified initially and 4 ADditional studies found in the references, one study that had to be removed as a duplicate. Due to missing main aspects of this review (no clinical population, no EEG, no deep learning) 92 studies were discarded at the first drop-out stage. After full text examinations, 20 more studies were removed from further consideration at a second drop-out stage because they were missing information on the main outcomes relevant to the review. Finally, a total of 30 studies are included for this review (see Fig. 1).

Table 1

| ICD-10 F-code | Search terms | Number of identified studies | Studies meeting criteria | Additional sources | Total Included | Comments |
|---------------|--------------|-----------------------------|-------------------------|-------------------|---------------|----------|
| F0            | dementia, alzheimer, delir, mild cognitive impairment | 11 | 3 | 0 | 3 | parkinson excluded |
| F0            | dementia, alzheimer, delir, mild cognitive impairment | 9 | 2 | 0 | 2 | 1 additional overlap with F0 |
| F2            | schizophrenia, delusional, schizotype, psychotic, schizoaffective | 13 | 3 | 2 | 5 | |
| F3            | mania, bipolar, unipolar, depression, affective | 56 | 9 | 1 | 10 | |
| F4            | phobia, anxiety, obsessive compulsive disorder, adjustment disorder, posttraumatic disorder, dissociative disorder, somatiform disorder, neurotic | 2 | 2 | 0 | 2 | |
| F5            | eating disorder, sleeping disorder, sexual disorder | 37 | 3 | 0 | 3 | no emotion processing studies |
| F6            | personality disorder, borderline disorder | 0 | 0 | 0 | 0 | |
| F7            | intellectual disabilities | 4 | 0 | 0 | 0 | |
| F8            | autism, asperger syndrome, developmental disorder | 6 | 4 | 1 | 5 | |
| F9            | ADHD, tic | 138 | 26 | 4 | 30 | |

Total 138 26 4 30
3.1. Clinical domain

3.1.1. Diagnosis

Most of the studies included in this review analysed data from patients suffering from depression (n = 10, Table 1). Studies on ADHD and schizophrenia are the next most frequent (both n = 5, Table 1). Fewer studies analysed dementia and sleeping disorders (both n = 3, Table 1). Finally, two studies were included in this review which analysed addictions and anxiety/phobias (n = 4, Table 1). We did not uncover any studies of personality disorders, intellectual disabilities, or autism which met our eligibility criteria in this review (see Table 1).

3.1.2. Validation of diagnosis

In the final corpus, only 15 out of 30 studies were included which contained a clear and reproducible diagnostic procedure based on the international classification systems (ICD-10 or DSM-V) or following international guidelines (Table 2). Of those 15 studies, three only partly reported on the diagnostic process they used and sometimes only described a “clinical examination” or a single questionnaire. In four studies, the diagnostic procedure was not clearly explained. In eight other studies, there was no diagnostic validation (Table 2).

3.1.3. Number of subjects

The studies in this review included a range of subjects from a high of 178 to a low of 8 (Balan et al., 2020). In total, nine studies included 100 or more subjects (Table 2), ten studies included between 50 but below 100 subjects, eight studies included from 25 to below 50 patients (including controls) and three studies included less than 25 subjects. Some of the studies were based on the same datasets: Acharya et al. (Acharya et al., 2018); Ay et al. (Ay et al., 2019); two studies by Chen et al. (Chen et al., 2019a; Chen et al., 2019b); and two studies by Li et al. (Li, La et al., 2019; Li et al., 2020).

3.1.4. Medication

Most studies (21 out of 30) did not report on medication or other treatment forms present in their sample. From nine studies that reported on medications (Table 2), four of them indicated that a washout phase had taken place before the EEG recording. One other study stated explicitly that there was no medication in place (Zhang, Li et al., 2020), while four other studies reported the presence of medication (Table 2).

3.1.5. Diagnosis, progress or response

Two studies used a DL approach in their investigations of the
progress of a mental disorder (Ahmed Aristizabal et al., 2021; Chu et al., 2018). Two other studies (Balan et al., 2019, 2020) used DL techniques for response prediction. All the other studies used DL for classification of diagnosis.

### 3.2. EEG data domain

#### 3.2.1. EEG-condition

All studies reviewed reported on the setting under which EEG was recorded (Table 3). In total, 17 studies used resting state EEG recordings, most of them with eyes closed, but some also used eyes open condition e. g. (Kim & Kim, 2018). Two of the resting state studies also used event related potentials. One study used sleep EEG data (Shahin et al., 2017), while nine other studies used event related potentials from mostly visual tasks. One study used continuous data from an auditory listening task, most of them with eyes closed, but some also used eyes open condition e. g. (Kim et al., 2018). Two other studies (Balan et al., 2019, 2020) used DL techniques for response prediction. All the other studies used DL for classification of diagnosis.

### 3.2.2. Sampling rate

Sampling frequency was reported in 28 out of 30 studies (Table 3) while two studies left the sampling frequency unclear (Balan et al., 2020). Several studies reported a downsampling before analysis (n = 3) or only reported on the sampling frequency of the data used for DL models (n = 1, (Mumtaz & Qayyum, 2019)). All the other studies reported that the sampling frequencies were the same as for the Deep Learning. While three studies used a sampling rate of 1000 Hz, seven used 500 Hz or 512 Hz, 13 used 250 Hz or 256 Hz, two used 200 Hz, and three used 128 Hz or 125 Hz.

#### 3.2.3. EEG channel numbers

In all 30 studies included in this review, information on the number of channels of the EEG device was present (Table 3). In seven of those studies, however, the number of channels used for recording was larger than the number of channels used for analysis. Two studies only used only 16 out of the 128 (Li, Zhang et al., 2019) or 11 out of the 64 available channels (Zhang et al., 2020) while one other study used only data from one channel out of the 61 available EEG channels (Kim et al., 2018). The following list display the number of channels that were used for DL: Two studies used 128 channels, four studies used between 50 and below 100 channels, five studies used below 50 and above 25 channels, seven studies used 19 channels (representing the most common used clinical channel number), three more studies used 16 channels, one with eleven channels, one with seven channels, two with three channels, four studies only used two channels and one used only one channel. Only eleven studies reported on the references they used to support their approaches. Three studies explicitly reported on the recording of additional channels from such non-EEG sources as electrocardiograms or electrooculograms.
et al., 2020). Most often, authors reported using independent component analysis (ICA) for artefact reduction (e.g. eye artefacts or heartbeat artefacts, n = 7 studies). Other studies used techniques involving gradient- or amplitude-thresholds (six studies) as well as manual artefact selection (n = 5 studies).

3.3. Deep learning domain

3.3.1. Feature engineering, input features and input dimensions

All 30 studies changed the format of the EEG data by splitting the EEG into shorter time frames before performing further feature engineering. Seven of those studies used a combination of time and frequency domain features with three-dimensional data (Table 3). Three of those studies applied wavelet transformations. Two studies used topographic maps which contained time domain and spatial features to yield three-dimensional data. One study used spectral topographic maps (Bi & Wang, 2019) which contained frequency domain features and spatial features (three-dimensional data). Although eleven studies only used time domain data (two-dimensional: time and channel), ten of those used some type of frequency filter. Moghaddari et al. created images in

Table 3

| Author                        | Task                        | EEG channels for DL input | Data sampling [Hz] | Preprocessing | Artifact handling |
|-------------------------------|-----------------------------|---------------------------|--------------------|---------------|-------------------|
| Acharya et al. (2018)         | Resting EEG                 | 2 bipolar channels 30     | 256                | Z-normalization | Manually          |
| Ahmad Aristizabal et al. (2021) | Auditory oddball            | 30                        | 500                | Notch filter   | Baseline correction, amplitude scaling |
| Ay et al. (2019)              | Resting EEG                 | 2 bipolar channels 32     | 256                | Notch filter   | Manually          |
| Balan et al. (2019)           | DEAP Database 40 min video  | 32                        | 512                | BPF 4–45 Hz, epoch to 60 s | Blind source separation |
| Balan et al. (2020)           | Virtual Reality Exposure Therapy | 8                      | ?                  | BPF, epoch to 1 s | The last correct value |
| Bi and Wang (2019)            | Resting EEG                 | 64                        | 500                | Epoching to 0.5 s segments | ? |
| Cai et al. (2016)             | Sound listening             | 3                         | ?                  | BPF 0.5–50Hz   | Wavelet transform |
| Chen et al. (2019a, 2019b)    | Resting EEG                 | 128, 62 used for analysis | 1000               | BPF 0.1–100 Hz, epoch to 4 s | ICA, voltage threshold |
| Chen et al. (2019a, 2019b)    | Resting EEG                 | 128, 32 used for analysis | 1000               | BPF 0.5–40 Hz, epoch to 4 s | ICA, voltage threshold |
| Chu et al. (2018)             | Resting EEG                 | 64                        | 1000               | BPF 0.01–50 Hz | Eye artefact removal |
| Dubreuil-Vall et al. (2020)   | Flanker Task and resting EEG | 7                        | 500                | BPF 1–20 Hz   | ICA               |
| Kim et al. (2018)             | Visual ERP                  | 61, 1 used for analysis   | 256                | No            | Delta encoding method |
| Kim & Kim (2018)              | Resting EEG                 | 32                        | 500                | BPF 1–55 Hz, epoch to 2 s, power normalization | ? |
| Kwon et al. (2019)            | Resting EEG                 | 19                        | 256                | ?             | ? |
| Li, La et al., Li, Zhang et al. (2019) | Emotional face task | 128, 16 used for analysis | 250                | BPF 0.5–40 Hz | ICA               |
| Li, Li et al., Li, Zhang et al. (2019) | Emotional face task | 128                      | 250                | BPF 1–40 Hz, epoch to 1 s | ICA |
| Li et al. (2020)              | Continuous mental task      | 128, 256                  | 250                | BPF 0.5–70 Hz | No                 |
| Moghaddari et al. (2020)      | Emotional face task         | 19, 512                   | 200                | Cascade filters (4–40 high-lowpass) | Artifact Subspace Reconstruction |
| Morabito et al. (2017)        | Resting EEG                 | 19                        | 256                | BPF 0.5–70 Hz, notch filter | Manually |
| Mumtaz and Qayyum (2019)      | Resting EEG                 | 19                        | 256                | No            | Multiple source eye correction |
| Oh et al. (2019)              | Resting EEG                 | 19                        | 250                | Epoch to 25 s, normalized with Z-score | ? |
| Phang et al. (2020)           | Resting EEG                 | 16                        | 128                | No            | ? |
| Shahin et al. (2017)          | Sleep EEG                   | 2 (C3, C4)                | 200                | BPF 0.5–40 Hz | Voltage threshold |
| Shalbaf et al. (2020)         | Resting EEG                 | 19                        | 250                | BPF 0.5–45 Hz, epoch to 5 s | ? |
| Skorucak et al. (2020)        | Maintenance of wakefulness test | 7, 2 used for analysis   | 200                | Cascade filters (0.3–70 high-lowpass) | ECG extraction |
| Uyulan et al. (2021)          | Resting EEG                 | 19                        | 125                | No            | ICA               |
| Vahid et al. (2019)           | ERP interval timing task    | 60, 56 used for analysis  | 500                | BPF 0.5–20 Hz, CSD | ICA, voltage threshold |
| Xie et al. (2020)             | Resting EEG and Stroop tasks | 31                      | 500                | BPF 0.5–100 Hz, notch filter | Voltage threshold, manually |
| Zhang, Silva et al. (2020); Zhang, Li et al. (2020) | Resting EEG | 3                        | 250                | BPF 1–40 Hz, epoch to 4 s | Wavelet and Kalman algorithm |
| Zhang, Silva et al. (2020); Zhang, Li et al. (2020) | Object recognition tasks | 64, 11 used for analysis | 256                | Value normalization to [0,1] | Voltage threshold |

Notes: 1Please refer to Balan et al. (2020) for technical information. Abbreviations: BPF, bandpass filter; ICA, independent Component Analysis; ?, No information/unclear.

3.2.4. Preprocessing

In the corpus of studies under review, bandpass filtering of the EEG signal was the most common technique for preprocessing. In total, 17 out of 30 studies (Table 3) used bandpass filters or a similar technique. Four studies did not preprocess their data (Kim et al., 2018; Mumtaz & Qayyum, 2019; Phang et al., 2020; Uyulan et al., 2021). Other preprocessing steps included normalization techniques, e.g. z-normalization (Acharya et al., 2018) or value normalization (Zhang, Silva et al., 2020). The preprocessing steps used in one study were not clearly described (Kwon et al., 2019).
which red-green-blue (RGB) channels are replaced by theta, alpha, and beta + low gamma waves. Three studies used only frequency domain features, two of which were two-dimensional and one was one-dimensional. One study used frequency and spatial features by adding a spectral topographic map to yield three-dimensional data. Further, four others used such connectivity features as phase lag or coherence between channel. Of those four, three of them used three-dimensional data while one other used one-dimensional data. Of the three studies that adopted a one-dimensional vector, one of them (Kim et al., 2018) used Wave2vec to reduce dimensionality.

3.3.2. Network structure

In total, 21 studies used a CNN type with 16 of them using two-dimensional CNN, one using a three-dimensional CNN (Chen et al., 2019b) and five other studies using one-dimensional CNNs. Two studies combined a one-dimensional CNN with a two-dimensional CNN (Li, La et al., 2019; Phang et al., 2020). Further, four studies used only fully connected networks. Two studies (Ahmed Aristizabal et al., 2021; Ay et al., 2019) combined CNN with LSTM, and only one study (Skorucak et al., 2020) used LSTM alone (Skorucak et al., 2020). One study combined CNN with an attention network (Zhang, Li et al., 2020) in order to include patient information, such as sex. One study used the generative Deep Belief Network. Four studies included pretrained models in the neural network while two other studies used an existing, pretrained CNN for feature extraction and combined it with a SVM for classification (ReNet: [Shalbaf et al., 2020]), Mobilenet: (Zhang, Silva et al., 2020)).

3.3.3. Hidden layers

The number of hidden layers ranged from one (Xie et al., 2020) to 50 (Uyulan et al., 2021), reflecting the wide range from “shallow” to “deep” networks (Table 4).

3.3.4. Activation function

Most studies (14) used the ReLU activation function. Two used Leacky-ReLU, and three used ELU. Four used sigmoid, and two used hyperbolic tangent function. In six studies, we found no information about the activation function.

| Author | Data features | Reported independency in testing set/folds | Statistics | Results | Reproducibility |
|--------|---------------|------------------------------------------|------------|---------|-----------------|
| Acharya et al. (2018) | Time, Frequency, Spatial | X | 0.96 | poor |
| Aristizabal et al. (2020) | | ✓ | 0.73 | good |
| Ay et al. (2019) | | X | 0.99 | good |
| Baldan et al. (2019) | | X | 0.82 | good |
| Baldan et al. (2020) | | ✓ | 0.72 | good |
| Bi et al. (2019) | | X | 0.95 | poor |
| Cai et al. (2016) | | X | 0.78 | poor |
| Chen et al. (2019) | | ✓ | 0.90 | middle |
| Chen et al. (2019) | | X | 0.95 | middle |
| Chu et al. (2018) | | X | 0.99 | poor |
| Dubreuil-Vall et al. (2020) | | ✓ | 0.88 | good |
| Kim et al. (2018) | Not applicable | ✓ | 0.69 | low |
| Kim & Kim (2018) | | X | 0.73 | middle |
| Kwon et al. (2019) | | X | 0.88 | poor |
| Li et al. (2019) | | ✓ | 0.85 | middle |
| Li et al. (2019) | | ✓ | 0.86 | middle |
| Li et al. (2020) | | ✓ | 0.81 | middle |
| Moghaddari et al. (2020) | | X | 0.98 | poor |
| Morabito et al. (2017) | | ✓ | 0.89 | poor |
| Mumtaz and Qayyum (2019) | | X | 0.98 | good |
| Oh et al. (2019) | | ✓ | 0.81 | middle |
| Phang et al. (2019) | | X | 0.92 | good |
| Shahn et al. (2017) | | X | 0.92 | middle |
| Shalbaf et al. (2020) | | X | 0.99 | middle |
| Skorucak et al. (2020) | | ✓ | 0.98 | middle |
| Uyulan et al. (2020) | | X | 0.93 | good |
| Vahid et al. (2019) | | ✓ | 0.83 | good |
| Xie et al. (2020) | | X | 0.68 | poor |
| Zhang et al. (2020) | | X | 0.75 | poor |
| Zhang et al. (2020) | | X | 0.95 | good |

Notes: Input from domain was marked in light grey; Results reported as accuracy.

3.3.5. Optimisation algorithm

Most studies (16) used ADAM for optimization, three used NADAM, and two used vanilla stochastic gradient descent. Six studies did not report on the optimization method they employed.

3.3.6. Model selection and testing strategy

Five studies used a separate set for testing, 23 used a CV technique, and two did not report on the testing method they employed. Since all 30 studies used a type of data reformatting for splitting EEGs into shorter timeframes, the various networks were fed with different timeframes from the same subject. From the five studies that used a separate test set, only two reported that they were testing on timeframes from subjects that were not included in the training dataset. From the 23 studies that used CV for testing, only twelve reported creating folds with independent subjects. In total, seven studies did not report whether the testing was done on independent subjects. Regarding whether the test set was kept “unseen” during model selection, four of the five studies which used a separate test set approach reported model selection without using the test set. Of the 23 studies that used CV for testing, 14 used a model selection technique within every training fold that allows CV to be run only once. Two studies did CV without model selection (an approach which also allows CV to run only once). Only two studies reported using a model selection technique that did not allow to run CV only once, while seven studies did not report what kind of model selection technique they used. In summary, of the 30 analysed studies, only eleven reported using model selections procedures that yield a correct testing technique.

3.3.7. Backward modelling

Nine studies included descriptions of backward modelling, while two studies used the Deep Dream technique (Dubreuil-Vall et al., 2020; Zhang, Li et al., 2020). Others implemented saliency maps (Vahid et al., 2019) in combination with source estimation software or gradient-weighted class activation mapping (Chen et al., 2019b).

3.3.8. Additional statistics

Besides the application of DL-networks for EEG-data classification,
five studies used additional conventional statistics to verify their findings or to compare their results with the results obtained by using other methods.

3.3.9. Reproducibility

Regarding reproducibility, it was found that only ten studies showed good reproducibility, ten others showed middle reproducibility, and ten others showed poor reproducibility.

3.3.10. Results

The reported outcomes differed to a large extent (see Table 4). All of the studies reported accuracy measures of their results. These levels of accuracy achieved in the thirty studies ranged from 69 % to 99 %.

4. Discussion

In conducting this review, a systematic search for published studies which utilized learning-based methods to analyse EEG data derived from patients suffering from mental disorders listed in ICD-10 was performed. In contrast to the anticipated increase in DL driven studies in all fields of medicine, only a relatively small number of studies using EEG in the field of psychiatry was found. (Hinton, 2018; Naylor, 2018). From a clinical perspective, it was discovered that most of the identified studies did not include a careful assessment of mental disorders based on their associated clinical data, and that many of them lacked clear enough descriptions of the procedures they used for valid reproducibility. It was also found that the independence of the test set used for obtaining the final accuracy was often compromised. Many authors used different EEG data segments from the same subjects in the training, validating, and testing dataset. Thus, the accuracy achieved in some studies reflects the ability of DL networks to learn EEG features from individual subjects but may not perform well on data of new patients. Although some of the results were in themselves impressive, the successful use of DL techniques on EEG data in psychiatry will depend on the correct usage of DL testing procedures and the sufficient characterization of the clinical populations used as subjects.

Although deep learning-based approaches have gained attention in the field of EEG analysis (Craik et al., 2019; Roy et al., 2019), the corpus of studies that were identified on mental disorders was still rather small. Nevertheless, the number of available studies has been gradually increasing over the past two years and is expected to increase in the future. In this review, it was found that the largest number of DL based EEG studies addressed major depression, which is among the leading causes of years lived with disability (Murray et al., 2012). Most of these studies did not include any consideration of such highly prevalent conditions as anxiety or obsessive-compulsive disorder; these associated conditions were missing or clearly underrepresented in the identified EEG-DL corpus. Conversely, it was found that many of the studies considered schizophrenia (Saha et al., 2005), a mental disorder which historically has comprised a large part of psychiatric research but is less prevalent than other conditions.

We found it striking that almost half of the studies that were examined neglected to address the clinical aspects in their designs: precise descriptions of the diagnostic process were missing, and the results of questionnaires rarely given. Another example of failed reports of relevant features are the missing statements of used medications in many of the reviewed studies. Although not very specific, the EEG is a very sensitive instrument able to capture activity changes due to psychopharmacological interventions. A change of e.g. the alpha peak frequency can be tracked easily (sensitive), but the changes do not allow for specific pathognomonic interpretation (unspecific. e.g. the reasons for alpha band slowing might be intoxication or a structural process such as dementia). However, EEG has been used for decades in the context of psychopharmacology (Jobert et al., 2012) and has a high clinical value in the monitoring of such different drugs as lithium and clozapine (Fogarell, 2017). It is well-known that DL networks can learn and extract patterns from different features (Najafabadi et al., 2015), therefore the classification of features derived from a coarse clinical diagnostic process are useless and potentially dangerous if used in the wrong context. For instance, in many studies, it is not clear if the DL network learns to classify the diagnosis, or if it learns to classify confounding variables such as medications associated to that diagnosis. If the field of DL aims to achieve a high standard of clinical relevance, studies must be serious in the way they use data derived from patients. Combining the clinically relevant data with DL will be the key for success in a clinical context. To achieve more solid testing accuracy in delineating clinical populations for clinical usage in the future, the networks used will need to be trained on data where such other factors as medication have been excluded or at least have been carefully described.

Regarding the quality of EEG recordings, it must be stated that the montage has not been reported in all studies, thus preventing the possibility of a straightforward replication. The number of channels used for recording seemed sufficient in most studies, but several of them reduced the channel number when calculating the input into the networks. One reason for the success of DL in image classification is that the first layers of the network learn to extract the right features whereas the last layers classify the extracted features. In general, the winners of ImageNet competitions in recent years did not use dimensionality reduction nor feature engineering since it lowers the amount of available data. They trained the networks to identify the right features. With increasing amount of data, it is believed that the highest level of accuracy can be achieved by feature extraction using DL models instead of human feature engineering. Furthermore, most of the best ImageNet scores have been achieved by using some sort of data augmentation. It is recommended that EEG data scientists focus in the future more on data augmentation instead of on feature engineering and dimensionality reduction. A really noteworthy approach for increasing the information in the data can be found in two of the reviewed studies where information on channel localisation is included in the input vector for DL analysis (Chen et al., 2019b; Li, La et al., 2019). Still, Li et al. (Li, La et al., 2019) report that the inclusion of spatial information did not increase the accuracy of their approach. In future studies, however, also the input of additional clinical data might enrich the models, as done by (Zhang, Li et al., 2020), where the information on age and gender were included. These approaches can be used to increase the amount of available information instead of reducing it.

Though most of the studies that were examined reported the DL architectures and data structures, many of them failed to mention the rationale behind the chosen architecture. Since a reasonable comparison between the studies was not possible because of the heterogeneity of performed classification tasks, it is still largely unknown which network structure might be best for EEG classification tasks when analysing psychiatric disorders. Furthermore, most of the studies did not try different approaches but relied mainly on a previously chosen model. Only some of the authors reported on their exhaustive testing of different network models (Phang et al., 2020) or on the impact of such changes in their preprocessing steps as filtering had on their results (Li, La et al., 2019). It is important to mention that some studies used preexisting networks from other domains (e.g., image recognition tasks (Kwon et al., 2019; Shalbaf et al., 2020; Uyulan et al., 2021; Zhang, Silva et al., 2020)). Furthermore, some studies also used pretrained layers (Bi & Wang, 2019; Shalbaf et al., 2020; Zhang, Silva et al., 2020). This applied transfer learning might be an important key for the successful implementation of DL techniques in clinical contexts since it allows for the use of sophisticated networks from other domains without having to train them on large datasets.

It was surprising to discover how, in some studies, the correct separation of training and testing data had been compromised. Further, it is noteworthy that most studies had no separate testing dataset but performed a cross validation approach instead. While on the surface this is a good state of the art practice, a separate test set should be used in DL models to yield the highest level of validity in the results. The reason
that most studies did not use separate test sets is the low numbers of included subjects. The availability of only limited data prevented most studies from creating an additional test set. Still, even when studies reported on a separate test set, some of them used data from the same subject in their training and test sets (Table 4). Since EEGs have high intersubject variability and a high intrasubject stability, large DL networks will not learn the common features across subjects but will recall different subjects by identifying their individual EEG signature. Although such networks can achieve a high level of accuracy, they become useless in a clinical context when they are applied to totally new datasets from different subjects.

Since this review has focused on the clinical side, it should be noted that the features that a deep learning network uses for classification can be extracted and analysed, which could lead to significant contributions to a better general understanding of psychiatric disorders. It is possible that the hidden patterns inherent in mental disorders which DL is capable of revealing might shed light on some of the basic mechanisms that underly these disorders. Thus, it is important to understand that some of these studies tried to extract these patterns to increase and deepen our knowledge of the electrophysiological representations of the studied disorders. To achieve accurate results, however, the use of these sophisticated DL backward modelling techniques (such as gradient-weighted class activation mapping for example (see Chen et al., 2019b)) need to be very carefully implemented and the results even more carefully interpreted to be useful in a clinical setting.

Based on the results of our review, we have developed the following list of recommendations to make it easier for researchers to more effectively use DL models on psychiatric EEG data in future psychiatric research studies:

1. Use clear terminology (e.g., “epochs” in EEG research and in Deep Learning models have different meanings (Roy et al., 2019)).
2. Carefully describe the clinical sample and identify any possible confounding variables including the presence of medications.
3. Validate the diagnostic procedures used against international standards.
4. Follow and fully describe the EEG standards used for recording and processing (Jobert et al., 2012).
5. Question feature engineering and explore data augmentation.
6. Choose a clear model selection strategy and report on it. Be careful not to include information present in the test set for model selection and hyperparameter tuning. Therefore, testing should be done only once. For hyperparameter tuning when testing with CV, model selection should be performed anew in every fold. For example, nested CV could be used.
7. Ensure that the test data is independent from the training data. Therefore, use separate subjects for training and testing.
8. Ensure to identify confounding variables and balance them between test and training data.
9. Choose a reasonable score for reporting results (e.g., for imbalanced data, choose a score such as the F1-score that accounts for precision and recall).
10. Analyse and report the influence of different hyperparameters, different network architectures, and the application of the preprocessing steps.
11. Increase transparency by describing the methods and models in detail and make the code publicly available whenever possible.

5. Conclusion

The emergence of new analysis procedures and technique in EEG research can enhance the way researchers use clinical EEGs: It can support to extract information and patterns from electrophysiological timethes that had previously been hidden behind the complex structures of EEG recordings. However, since the promise of a technological transformation has been made for decades in psychiatric EEG research, researchers and clinicians need to be better informed about the many pitfalls that this systematic review identifies and describes. Based on the mistakes and the flaws that have been identified, future researchers can be better equipped in the future to work more effectively together to create clinically sound, meaningful, and valid DL-based studies.

Declaration of Competing Interest

CI declare that she has no material or financial interests related to the research described in this paper. MB and SO report that they are the founders of a deep learning-based start-up for EEG classification. MB and SO have no other competing interests to report.

Acknowledgement

None.

References

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adeli, H., & Subha, D. P. (2018). Automated EEG-based screening of depression using deep convolutional neural network. Computer Methods and Programs in Biomedicine, 161, 103–113. https://doi.org/10.1016/j.cmpb.2018.04.012
Ahmed Aristizabal, D., Fernando, T., Denman, S., Robinson, J. E., Sridharan, S., Johnston, P. J., … Fookes, E. (2021). Identification of children at risk of schizophrenia via deep learning and EEG responses. IEEE Journal of Biomedical and Health Informatics, 25(1), 69–76. https://doi.org/10.1109/JBHI.2020.2994238
Ay, B., Yildirim, O., Talo, M., Baloglu, U. B., Aydin, G., Puthankattil, S. D., & Acharya, U. R. (2019). Automated depression detection using deep representation and sequence learning with EEG signals. Journal of Medical Systems, 43(7), 205. https://doi.org/10.1007/s10916-019-1345-y
Balan, O., Moise, G., Moldoveanu, A., Leordeanu, M., & Moldoveanu, F. (2019). Fear level classification based on emotional dimensions and machine learning techniques. Sensors (Basel, Switzerland), 19(7). https://doi.org/10.3390/s19071738
Balan, O., Moise, G., Moldoveanu, A., Leordeanu, M., & Moldoveanu, F. (2020). An investigation of various machine and deep learning techniques applied in automatic fear level detection and acrophobia virtual therapy. Sensors (Basel, Switzerland), 20(2). https://doi.org/10.3390/s20020496
Berger, P. D. H. (1929). Über das Elektrenkephalogramm des Menschen. Archiv für Psychiatrie und Nervenkrankheiten, 87(1), 527–570. https://doi.org/10.1007/BF01797193
Bhugra, D., Tasman, A., Pathare, S., Priebe, S., Smith, S., Torous, J., … Ventriglio, A. (2017). The WPA-lancet psychiatry commission on the future of psychiatry. The Lancet Psychiatry, 4(10), 775–818. https://doi.org/10.1016/S2215-0366(17)30333-0
Bi, X., & Wang, H. (2019). Early Alzheimer’s disease diagnosis based on EEG spectral images using deep learning. Neural Networks: The Official Journal of the International Neural Network Society, 114, 119–135. https://doi.org/10.1016/j.neunet.2019.02.005
Boutros, N. N. (2015). Neuropsychiatric electrophysiology. Neuropsychiatric Electro physiology, 1(1), 4. https://doi.org/10.4138/138010-15100806
Gai, H., Sha, X., Han, X., Wei, S., & Hu, B. (2016). Pervasive EEG diagnosis of depression using deep belief network with three-electrodes EEG collector. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1239–1246. https://doi.org/10.1109/BIBM.2016.7822696
Chen, H., Song, Y., & Li, X. (2019a). A deep learning framework for identifying children with ADHD using an EEG-based brain network. Neurocomputing, 356, 83–96. https://doi.org/10.1016/j.neucom.2019.04.058. Scopus.
Chen, H., Song, Y., & Li, X. (2019b). Use of deep learning to detect personalized spatial-frequency abnormalities in EEGs of children with ADHD. Journal of Neural Engineering, 16(6), 066046. https://doi.org/10.1088/1741-2552/aaba3a
Chu, L., Qiu, R., Liu, H., Ling, Z., Zhang, T., & Wang, J. (2018). Individual recognition in schizophrenia using deep learning methods with random forest and voting classifiers: Insights from resting state EEG streams. ArXiv:1707.03467 [Cs] http://arxiv.org/abs/1707.03467
Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: A review. Journal of Neural Engineering, 16(3), 031001. https://doi.org/10.1088/1741-2552/ab0db9
Cutbberth, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. BMC Medicine, 11(1), 126. https://doi.org/10.1186/1741-7015-11-126
Dubrull-Vall, L., Ruffini, G., & Campodon, J. A. (2020). Deep learning convolutional neural networks discriminate adult ADHD from healthy individuals on the basis of event-related spectral EEG. Frontiers in Neuroscience, 14, 251. https://doi.org/10.3389/fnins.2020.00251
Eisenberg, L. (1987). Progress in neuropsychiatry. The Journal of Pediatrics, 51(3), 334–349. https://doi.org/10.1016/S0022-3476(57)80210-8
Faw, J. D., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Giorini, X., Donohue, B., Visentin, D., van den Driesche, G., Lakshminarayanan, B., Meyer, C., Mackender, F., Bouton, S., Ayoub, K., Chopra, R.,...
Hughes, J. R. (1996). A review of the usefulness of the standard EEG in psychiatry.

Itil, T. M. (1983). The discovery of antidepressant drugs by computer-analyzed human cerebral bio-electrical potentials (CEEG).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief networks.

Hinton, G. (2018). Deep learning.

Moghaddari, M., Lighvan, M. Z., & Danishvar, S. (2020). Diagnose ADHD disorder in children using convolutional neural network based on continuous mental task EEG.

Morabito, F. C., Campolo, M., Mammone, N., Francescatti, S., Tagliavini, F.,... & Conradi, J. (2016). Future of clinical EEG in psychiatric disorders: Shifting perspectives of promising baseline EEG biomarkers in major depressive disorder and attention deficit hyperactivity disorder.

Phang, C.-R., Noman, F., Hussain, H., Ting, C.-M., & Ombao, H. (2020). A multi-domain connectome convolutional neural network for identifying schizophrenia from EEG connectivity patterns. IEEE Journal of Biomedical and Health Informatics, 24(5), 1333–1343. https://doi.org/10.1109/JBHI.2020.2941222

Pogarell, O. (2017). EEG in der Psychiatrie. Das Neurophysiologie-Labor, 39(3), 116–128. https://doi.org/10.1016/j.neulab.2017.06.012

Rajpurkar, P., Yang, J., Das, N., Vale, V., Keller, A. S., Irvin, J.,... & Smoller, J. W. (2018). Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: A prespecified secondary analysis of a randomized clinical trial. JAMA Network Open, 3(6), Article e206653. https://doi.org/10.1001/jamanetworkopen.2020.6525

Rolle, C. E., Fonzo, G. A., Wu, T., Roll, R., Jha, M. K., Cooper, C.,... & Ekin, A. (2020). Cortical connectivity moderators of antidepressant vs placebo treatment response in major depressive disorder: Secondary analysis of a randomized clinical trial. JAMA, 320(3), 397–408. https://doi.org/10.1001/jama.2019.28962

Roy, Y., Banville, H., Albuquerque, I., Filmgot, A., Lack, T. F., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: A systematic review. Journal of Neural Engineering, 16(5), 051001. https://doi.org/10.1088/1748-0333/ab39a0

Smoller, J. W., Andrews-Gardner, A., Edenberg, H. J., Farone, S. V., Glatt, S. J.,& Kendler, K. S. (2019). Psychiatric genetics and the structure of psychopathology. Molecular Psychiatry, 24(3), 409–420. https://doi.org/10.1038/s41380-019-0104-0

Tye, K. M., & Deisseroth, K. (2012). Optogenetic investigation of neural circuits underlying brain disorders. Nature Reviews Neuroscience, 13(4), 251–266. https://doi.org/10.1038/nrn3171

Uyukan, C., Erzgul, T. T., Unalb, H., Cebi, M., Sayar, G. H., Neshad Aazd, M., & Tarhan, N. (2021). Major depressive disorder classification based on different convolutional neural networks models: Developing learning approach. Clinical EEG and Neurosci, 52(1), 38–51. https://doi.org/10.1177/1550059420916634, 10.1177/1550059420916634

Vahid, A., Bluschke, A., Roessner, V., Stober, S., & Beste, C. (2019). Deep learning based on event-related EEG differentiates children with ADHD from healthy controls. Journal of Clinical Medicine, 8(7). https://doi.org/10.3390/jcm8071055

Widowicz, D., Bilge, M. T., Montana, R., Chang, W., Rodriguez, C. I., Deckenber, N.,... & Neufeldt, C. B. (2019). Electroencephalographic biomarkers for treatment response prediction in major depressive illness: A meta-analysis. The American Journal of Psychiatry, 176(1), 44–56. https://doi.org/10.1176/appi.ajp.2018.17121538

World Health Organization. (1993). The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research. World Health Organization. https://apps.who.int/classifications/ICD10/en/

Wu, W., Zhang, T., Jiang, J., Lucas, M. V., Fonzo, G. A., Rolle, C. E.,... & Zipursky, R. B., Meyer, J. H., & Verhoeff, N. P. (2007). PET and SPECT imaging in psychiatric disorders. The ICD-10 classification of mental and behavioral disorders. World Health Organization, 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010.

Zhang, H., Huang, S. H. F., Ohata, E. A., Medeiros, G. A., & Rebeca, K. H. (2002). Bi-directional approach based on overlap between alcohols pre-disposition classification via EEG signals. Frontiers in Human Neuroscience, 14, 365. https://doi.org/10.3389/fnhum.2020.00365

Zipsky, R. B., Meyer, J. H., & Verhoeff, N. P. (2007). PET and SPECT imaging in psychiatric disorders. Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie, 52(3), 146–157. https://doi.org/10.1177/07067437073052003